ExPlain 3.0 manual

explaining gene expression data

November 2, 2009

Copyright © 2009 BIOBASE GmbH

Contents

I User manual 7

1 Main components of the ExPlain user interface 9
- 1.1 The project tree . . . 10
  - 1.1.1 Search tab on the tree frame . . . 11
  - 1.1.2 Select tab on the tree frame . . . 11
- 1.2 The menu and dialogs . . . 13
- 1.3 Layouts . . . 15
- 1.4 The process monitor . . . 15
- 1.5 The workspace . . . 16
  - 1.5.1 Renaming project tree nodes . . . 16
  - 1.5.2 Item origin information . . . 16
  - 1.5.3 Filtering table . . . 17
  - 1.5.4 Adjusting the number of visible table rows and page navigation . . . 19
  - 1.5.5 Sorting, renaming and showing/hiding table columns . . . 20
  - 1.5.6 Customization of the column content display . . . 21
  - 1.5.7 Selecting rows and further actions . . . 21
  - 1.5.8 Data export . . . 22
  - 1.5.9 Interface preferences . . . 22

2 Analytical workflows and Wizard mode 25
- 2.1 Workflow . . . 25
- 2.2 Data loading . . . 25
  - 2.2.1 Gene set . . . 25
  - 2.2.2 List with fold change . . . 27
  - 2.2.3 Microarray series . . . 28
- 2.3 Workflow mode . . . 28

3 Gene sets 33
- 3.1 Loading data into the ExPlain system . . . 33
  - 3.1.1 Load gene set dialog . . . 33
    - 3.1.1.1 Supported species . . . 34
    - 3.1.1.2 Supported file formats . . . 34
  - 3.1.2 Import options of geneset loading dialog . . . 34
  - 3.1.3 New gene set dialog . . . 37
  - 3.1.4 Import data . . . 37
- 3.2 Representation of the gene set data . . . 39
- 3.3 Recombining gene sets . . . 40
  - 3.3.1 Filter gene set by condition . . . 41
  - 3.3.2 Filter gene set by other gene sets . . . 42
  - 3.3.3 Join two gene sets . . . 42
  - 3.3.4 Extracting unique genes . . . 42
  - 3.3.5 Extract Up/Down/Non-change . . . 43
  - 3.3.6 Extracting random subsets . . . 45
- 3.4 Adding columns to the gene set . . . 45
  - 3.4.1 Calculation from an existing numerical column . . . 45
  - 3.4.2 Linking a column from another data set . . . 46
  - 3.4.3 Adding a column with system annotations . . . 47
- 3.5 Exporting gene sets to BKL . . . 48
  - 3.5.1 Exporting selected genes as BKL search result . . . 48
- 3.6 Statistics calculator . . . 49

4 The Functional classification **51**
4.1 The Functional classification analysis . . . 51
4.1.1 How to run classificaiton analysis . . . 51
4.1.2 Functional Analysis categories . . . 52
4.1.3 Functional analysis dialog window . . . 53
4.2 Example analysis with BKL curated GO annotation . . . 53
4.3 Functional Analysis output tables . . . 54
4.3.1 Expression, BKL manual curation . . . 55
4.3.2 Gene Ontology analysis (public and BKL curated) . . . 55
4.3.3 Organ/tissue analysis output . . . 56
4.3.4 BKL Disease analyses output . . . 56
4.3.5 SwissProt analysis output . . . 56
4.3.6 Transcription Factor Classification analysis output . . . 56
4.3.7 TRANSPATH Molecule Classification analysis output . . . 57
4.3.8 Whole subsets from the tree analysis output . . . 57
4.4 Functional Analysis with reaction pathways . . . 57
4.4.1 Functional Analysis with TRANSPATH pathways . . . 58
4.4.2 Pathways results . . . 58
4.4.3 Functional Analysis with user-defined interaction pathways . . . 58
4.5 Functional Analysis summary . . . 59
4.6 Functional Analysis algorithm . . . 60
4.7 Gene Set Enrichment Analysis . . . 62
4.7.1 The GSEA interface . . . 62
4.7.2 GSEA example . . . 63
4.7.3 Enrichment Analysis theoretical basis . . . 63

5 Transcription factor site search **67**
5.1 Searching for sites in promoters . . . 67
5.1.1 How to run Match . . . 67
5.1.2 Sites search dialog window . . . 68
5.1.3 The Match output . . . 69
5.1.4 Optimization mode . . . 70
5.1.5 P-Match: combine pattern and matrix search . . . 71
5.1.6 Searching for sites in promoters using phylogenetic filtering . . . 72
5.1.7 Pattern based search . . . 73
5.1.8 De-novo motifs identification . . . 74
5.1.9 Filtering site search results using intervals . . . 74
5.2 Detailed graphical report of matrix distribution in promoters . . . 76
5.2.1 Matrix legend . . . 76
5.2.2 Promoter table . . . 77
5.2.3 Hiding matrices from view . . . 77
5.2.4 Detailed promoter view . . . 78
5.3 Summary set of several site search results . . . 79
5.4 Sites search theoretical background . . . 80
5.4.1 The TRANSPro database . . . 80
5.4.2 Computational definition of transcription start sites . . . 80
5.4.3 Cut-off values . . . 81
5.4.4 Calculation of theoretical P-values for TFBS predictions . . . 82
5.4.5 Sites search optimization with F-Match algorithm . . . 83
5.4.6 The P-Match algorithm . . . 84
5.4.7 The Patch algorithm . . . 85
5.4.8 The Seeder algorithm . . . 85

6 Composite module analysis and models 87
6.1 The CMA interface in ExPlain . . . 87
6.1.1 The CMA dialog window . . . 87
6.1.2 The CMA advanced dialog window . . . 88
6.1.3 The CMA output . . . 90
6.2 Composite modules on promoters . . . 94
6.3 Predefined parameters of CMA . . . 95
6.4 Composite element models . . . 95
6.5 Model editor . . . 96
6.6 Classifying promoters . . . 98
6.7 Obtaining interactions between Transcription factors and their target genes . . . 98
6.8 CMA - Composite Module Analyst – Background information . . . 99
6.8.1 CMA promoter models . . . 101
6.8.2 Model construction . . . 102

7 Molecular networks analysis 103
7.1 Network key node analysis . . . 103
7.1.1 Key nodes dialog window . . . 103
7.1.2 Results of key node analysis . . . 104
7.1.3 Summary of key node results . . . 105
7.1.4 Key node search algorithm . . . 107
7.2 Network cluster analysis . . . 107
7.2.1 Cluster dialog window . . . 107
7.2.2 Network cluster analysis results . . . 108
7.3 Network visualization . . . 109
7.3.1 Adding expression data to the network . . . 111
7.4 User defined interactions . . . 112
7.4.1 Loading of user defined interactions . . . 112
7.4.2 Interaction representation . . . 112
7.4.3 Joining interaction profiles . . . 112
7.5 The BKL database . . . 112

8 Profiles 117
8.1 Loading profiles . . . 117
8.2 Creating a new profile . . . 118
8.2.1 New profile dialog . . . 118
8.2.2 Example of profile creation . . . 118
8.2.3 Profile modification . . . 120
8.3 Creating profiles from gene sets . . . 120
8.4 Profile representation in the result table . . . 121
8.5 Profile menu options . . . 121
8.5.1 Create profile from selection . . . 122
8.5.2 Create gene set using selected matrices . . . 123
8.5.3 Extend set of selected matrices by all homologous matrices . . . 123
8.5.4 Change cut-offs . . . 123
8.5.5 Merge several profiles into one profile . . . 123
8.6 User matrices . . . 124
8.6.1 Creating new matrix . . . 125
8.6.2 Representation of the user matrix . . . 125
8.6.3 Changing factors associated with matrix . . . 126
8.6.4 Importing user matrices from TRANSFAC . . . 127

9 Genome intervals (ChIP-chip, TFBS, ChIP-Seq, Tiling arrays) 129
9.1 Loading of genome intervals . . . 129
9.1.1 Loading of genome intervals data from BED file . . . 129
9.1.2 Loading of genome intervals from CHP/BAR file . . . 129
9.1.3 Loading of genome intervals from Illumina BED file . . . 130
9.1.4 Intervals representation . . . 130
9.2 Recombining intervals . . . 131

9.2.1 Filtering intervals by conditions . . . 131
9.2.2 Filtering intervals by gene set . . . 132
9.3 Filtering of MATCH results using intervals . . . 132
9.4 Theoretical background of Illumina BED files processing . . . 133

10 Sequences **135**
10.1 Loading sequence data into ExPlain . . . 135
10.1.1 Load sequence data from a file . . . 135
10.1.2 Supported file formats . . . 135
10.1.3 Copy/Paste sequence data . . . 136
10.2 Sequences in ExPlain: example . . . 137

11 Statistical Analysis of Microarray Data **139**
11.1 Loading of CEL files and assignment of factor-level information . . . 139
11.1.1 Factor level assignment . . . 140
11.2 Low level analysis . . . 140
11.3 Quality control . . . 142
11.4 High level analysis . . . 145
11.4.1 Meta-analysis . . . 146
11.4.2 Statistical Analyses of the gene expression data . . . 148
11.4.3 The Rank Product algorithm . . . 150
11.5 CRC clustering . . . 151
11.5.1 Algorithm details of CRC clustering . . . 152

12 miRNA analysis **155**
12.1 Identification of miRNA targets . . . 155

13 Reports **157**
13.1 Report generation . . . 157
13.2 Graph report generation . . . 157
13.2.1 Generating a graph of the value distribution of a gene set . . . 157
13.2.2 Graphical report on intervals . . . 157

II Appendix **161**

14 References **163**

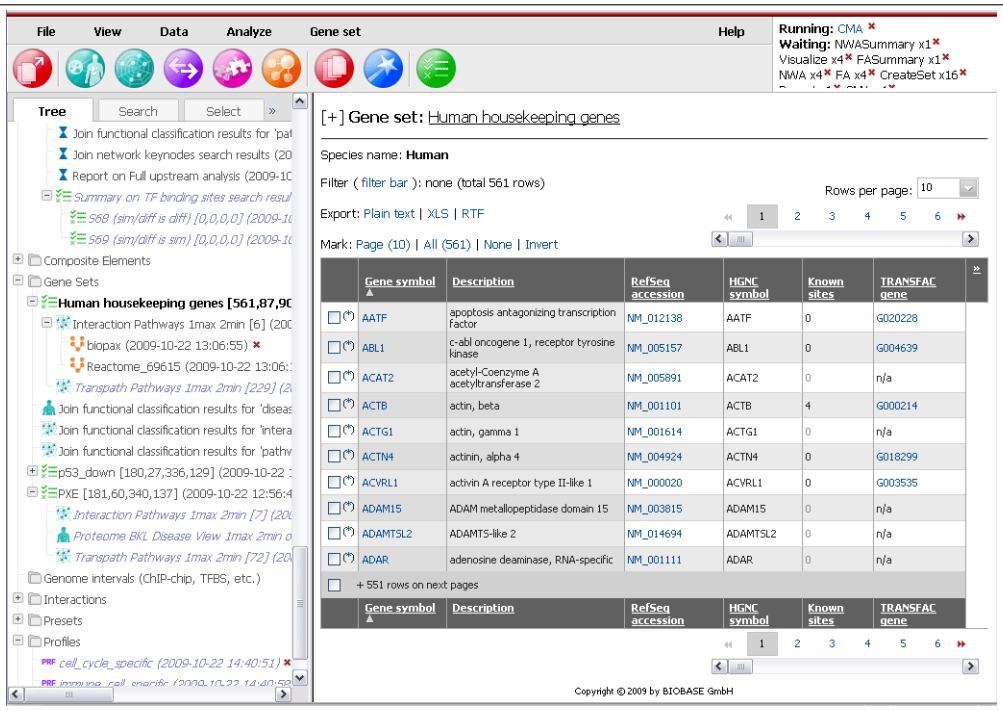
Part I

User manual

Chapter 1

Main components of the ExPlain user interface

Figure 1.1 The ExPlain user interface



The ExPlain interface has six major components:

Project tree The project tree gives access to all data sets and analysis results generated by ExPlain

Workspace (Input/output frame) The workspace displays actual data, such as gene sets, genomic intervals and the results of all analyses.

Menu The menu provides access to all analysis, data manipulation options and actions. If some action needs additional information from the user it will create a dialog providing functionality for specific tasks.

Toolbar The toolbar gives fast access to the most important items of the menu.

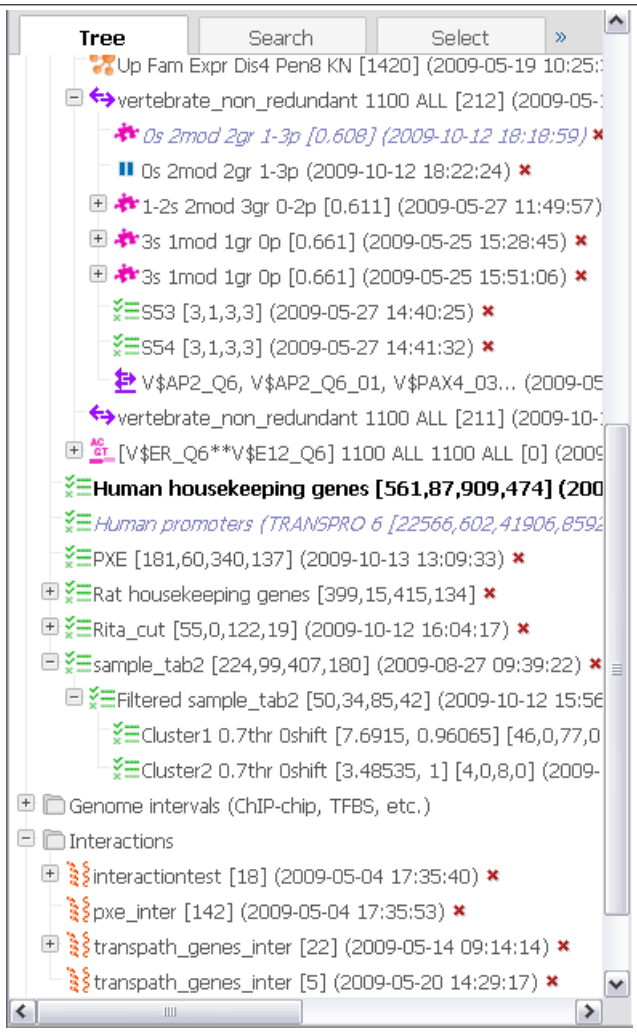
Dialog window All dialogs open in a separate window and provide functionality for specific tasks.

Process monitor The process monitor displays information about running, waiting or finished processor jobs.

1.1 The project tree

The project tree is the central access point for all data such as gene or protein data sets, derived subsets (see e.g. Chapter 3, «Gene sets»), analysis results, predefined or user-defined PWM profiles, promoter models or interaction profiles. When ExPlain is started in the browser, all data related to the corresponding user is compiled from the ExPlain database into an individual tree representation. The user's name is displayed in bold at the top of the tree.

Figure 1.2 The project tree



The project tree arranges all data into a hierarchy of nodes. The tree nodes are marked by different icons depending on their type. A complete list of icons and descriptions of corresponding data types is given in Table 1.1. There is typically an active node (displayed in bold in the tree), whose data is currently displayed in the Workspace (see also Figure 1.2). All nodes referring to analysis results or any data subset created during the analysis are children of the starting data node, provided as input for the analysis. This makes it easy to trace the workflow of a project. All nodes that have children nodes are prefixed with a ⊟ or ⊞ icon while they are in collapsed or expanded state respectively, and clicking on either icon expands or collapses the tree accordingly.

System folders (such as Composite Elements, Gene Sets, Genome intervals) are always present in the tree. The folder "Gene Sets" contains several data sets provided by Biobase, as well as input data and results of your analyses. Other system folders contain additional data sets (shown below) which can be used in various analyses.

Composite Regulatory Element models for CMA[2] compiled from TRANSCompel composite elements [1];

TRANSFAC profiles for binding site analysis [1];

User defined interactions for Key Node and Cluster analyses;

Genome intervals from ChIP-on-chip or ChIP-seq experiments for filtering of the site analyses results;

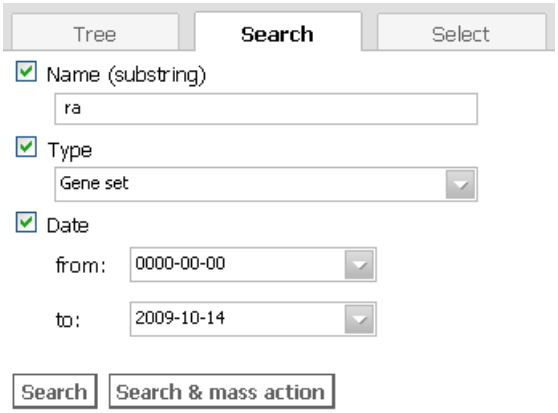
Parameter presets for data loading and CMA analyses.

New user folders can be created inside system foldersusing the command "New folder..." from the "File" menu.

1.1.1 Search tab on the tree frame

The "Search" tab can be used to set constrains on the node name, its type and time of creation. Press the Search button to simply highlight appropriate nodes or the Search & mass action button to higlight the nodes and switch to the "Select" tab in one step.

Figure 1.3 Search tab of the project tree



1.1.2 Select tab on the tree frame

The "Select" tab is useful for manipulation of project tree nodes. It gives you the opportunity to manipulate several tree nodes at once. Choose several nodes, select the appropriate action and press the OK button.

A convenient way for range selection and deselection provided by the tree select tab is to use the [Shift] key. The status of the last modified checkbox can be transferred to a whole range of rows by pressing the [Shift] key while selecting the top or bottom node of the desired range.

Table 1.1 Project tree icons

Icon	Description
	General icons
	Expands a node
	Collapses a node
	Expands all nodes
	Collapses all nodes
	Deletes a node from the tree
	Shows date of node creation
	Hides date of node creation
	A system folder
	A user created folder
	A process about to be launched
	An active process
	A paused process
	A postponed process
	An analysis process that could not be completed
	Data types icons
	A data set or subset
	Results from the Function analysis
	Results from the Function analysis in ExPlain Plant
	Results from the Pathways analysis
	A PWM profile
	Results of a binding site search analysis
	Motif search result
	Saved report page of a binding site search analysis
	Results of a promoter model construction with CMA
	Model created by a user in the model editor or provided by TRANSCompel
	A promoter model computed by CMA or results of a promoter model search in CMAcalc analysis
	Summary report page
	Graph report page
	Results of a pathway network analysis
	Visualization of a signalling network
	Interaction profile
	Genome intervals set
	Enrichment analysis results
	User-created weight matrix
	User defined data loading schema
	Predefined set of CMA parameters
	User defined set of CMA parameters
	A set of CEL-files in one archive
	A single CEL-file
	Quality check options
	Quality check filter
	T-test, Wilcoxon, GLM, ANOVA, Empirical Bayes, Rank-product test
	miRNA search result

Figure 1.4 Tree with the highlighted search results

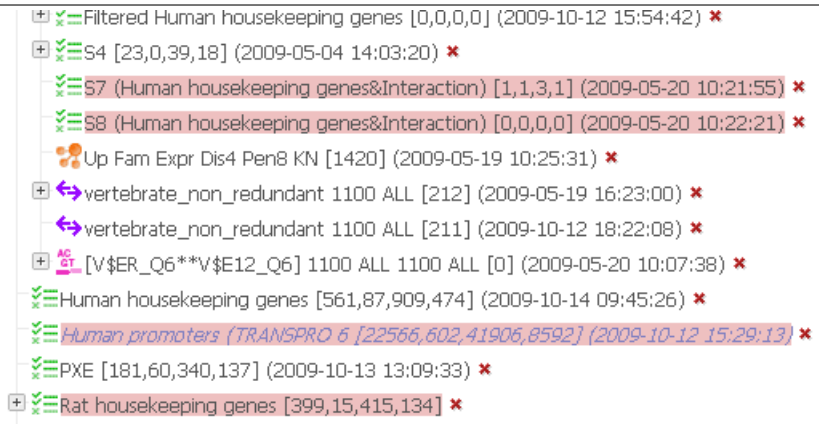


Figure 1.5 Select tab on the tree frame

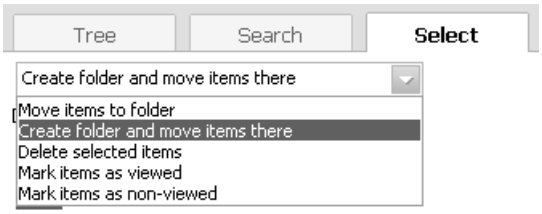
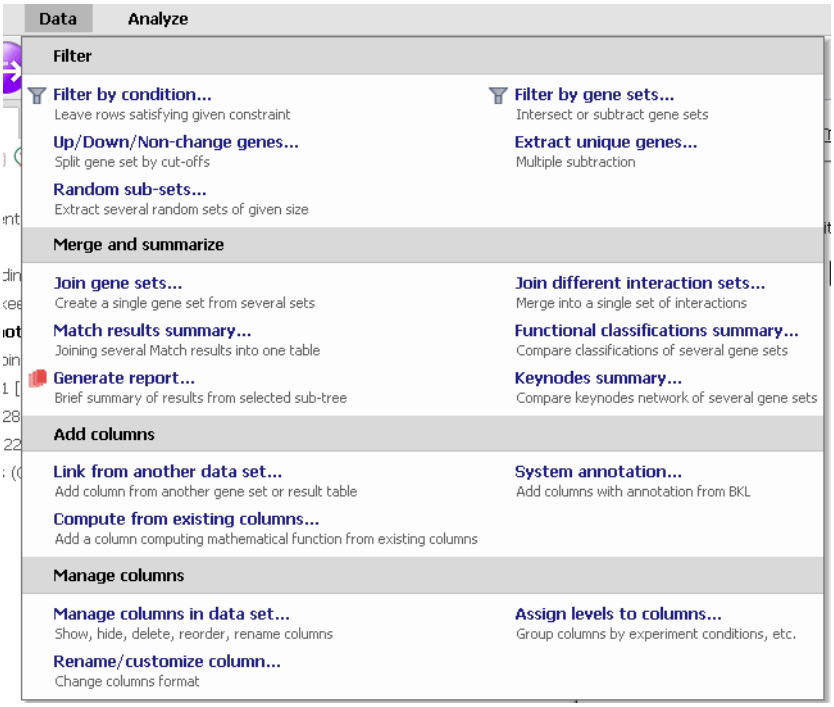


Figure 1.6 The menu


Data
Analyze
Filter
Filter by condition...
Leave rows satisfying given constraint
Filter by gene sets...
Intersect or subtract gene sets
Up/Down/Non-change genes...
Split gene set by cut-offs
Extract unique genes...
Multiple subtraction
Random sub-sets...
Extract several random sets of given size
Merge and summarize
Join gene sets...
Create a single gene set from several sets
Join different interaction sets...
Merge into a single set of interactions
Match results summary...
Joining several Match results into one table
Functional classifications summary...
Compare classifications of several gene sets
Generate report...
Brief summary of results from selected sub-tree
Keynodes summary...
Compare keynodes network of several gene sets
Add columns
Link from another data set...
Add column from another gene set or result table
System annotation...
Add columns with annotation from BKL
Compute from existing columns...
Add a column computing mathematical function from existing columns
Manage columns
Manage columns in data set...
Show, hide, delete, reorder, rename columns
Assign levels to columns...
Group columns by experiment conditions, etc.
Rename/customize column...
Change columns format

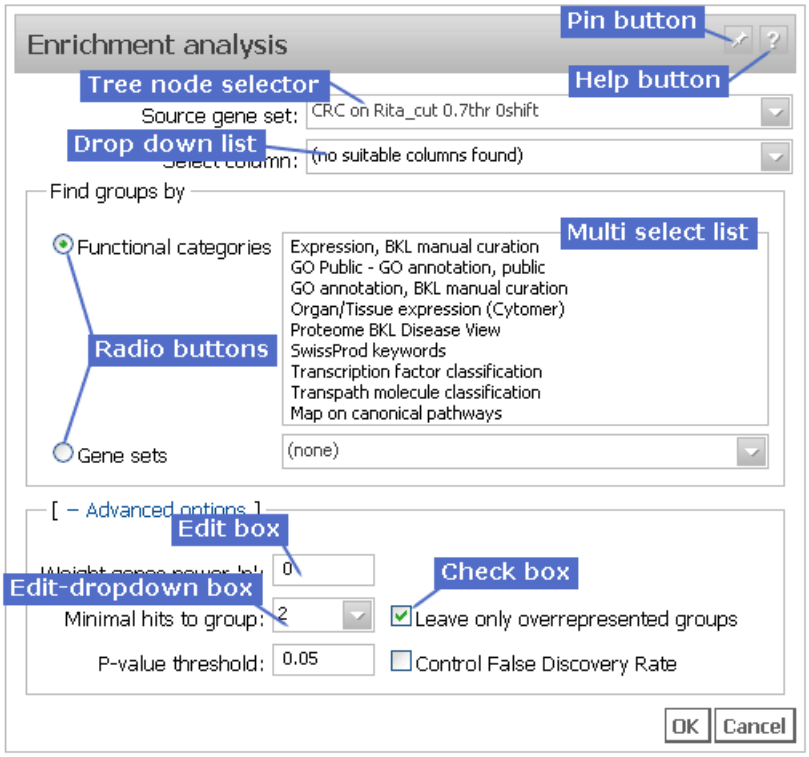
1.2 The menu and dialogs

The menu bar contains four constant items (File, View, Data and Analyze) and one variable item that differs depending on the selected tree node, containing actions specific for this node. Click on the menu item in the drop down to start a process or open the dialog window.

The dialog window contains several fields to configure program options and buttons to start an analysis or apply changes. The dialog window for the Enrichment analysis is shown in Figure 1.7.

With the check boxes, you can switch certain option on or off, with radiobuttons you can select one

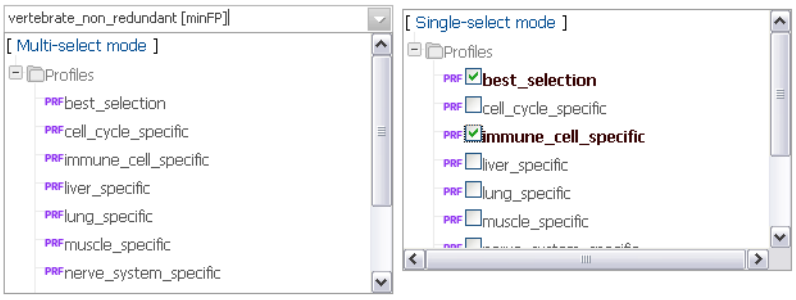
Figure 1.7 The dialog window exemplified by the Enrichment analysis



out of several options. Edit boxes are used to enter numerical or text values. Edit-dropdown boxes provide you with a list of suitable values, but you can also enter your own value manually.

The drop down list can be used to select one of the list items. One specific case of the drop down list is the tree node selector. By default this control shows only part of the tree that contains elements of a specific type. Nodes currently selected are shown in black, nodes that cannot be selected are shown in grey. This control can be in single-select or multi-select mode (see Figure 1.8). When both modes are available you can switch from one to another by using the [Multi-select mode] or [Single-select mode] links. In multiselect mode you can check several nodes to use them in some action or to start several processes, depending on the dialog. You can use the [Shift] key to select or deselect a range of nodes. If the first clicked node of the range is selected, then all other nodes in the range will be selected too, the same applies for deselection.

Figure 1.8 Tree node selection control in single-select and multi-select modes



The multiselect list provides you with the opportunity to select one or several items. Click on the line with the [Ctrl] key pressed to select several items, click using the [Shift] button to select or deselect a range of values, and single-click to select a single line and deselect all the others.

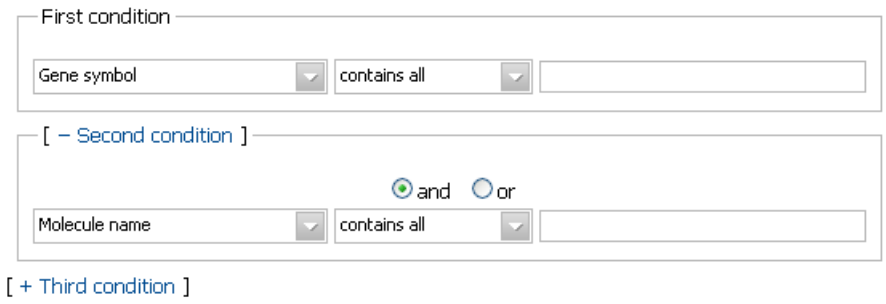
The "Help" button will lead you to the appropriate documentation chapter, describing data, analysis or dialog windows that were open at the moment you pressed the help button.

The dialog window will be closed when you press the OK button. To leave the window open, press the pin button, the picture will be changed to a . Press this button again to return to the original

window behavior.

In some dialogs additional options can be hidden in expanding blocks. To reveal these options click the link in brackets [+ Advanced options] . The "+" sign of the link will be changed to "-" for expanded options, as shown above. To collapse options back click the link again. On the figure below you can see an example of expanding blocks of a filtering dialog. The second filtering condition is expanded, and the third one is collapsed.

Figure 1.9 Expanding blocks



In some cases when one control is changed, other controls may become grey striped and the OK button is disabled (see the figure below). After ExPlain loads appropriate values, these controls will be available again.

Figure 1.10

1.3 Layouts

The ExPlain user interface provides several layouts to make work more convenient. You can switch between layouts using the "Layout" submenu in the "View" menu. The "Classic" layout is set by default. The "Compact" layout doesn't have the toolbar providing a little more vertical screen space for the data frame and may be useful if you want to view more table rows at once. "Flip panels" and "Compact flip" layouts are variants of "Classic" and "Compact" with the tree panel on the right side.

The "treeless" layout is somewhat different. In this mode the tree is hidden by default, leaving all of the horizontal screen space for your data, which might be useful when you have many columns. Instead of the tree, a breadcrumb bar is shown below the toolbar which displays the current item and all its parents allowing you to easily go up the tree. For the most part this is enough for navigation, though you still can see the whole tree by pressing the "Go to" button on the left side of breadcrumb bar. The only drawback to this mode is that tree functionality is limited: to search the tree or perform a mass action you'll have to switch layouts.

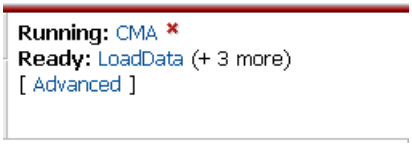
Figure 1.11 Breadcrumb bar

1.4 The process monitor

The process monitor provides an interface to the queue of ExPlain processes. The monitor window summarizes all running and waiting tasks as well as completed jobs whose results are waiting to be inspected. For further details about a running or waiting process, you can click on its label in the monitor window. Processes can be cancelled by clicking on the ✖ icon associated with a label. The process

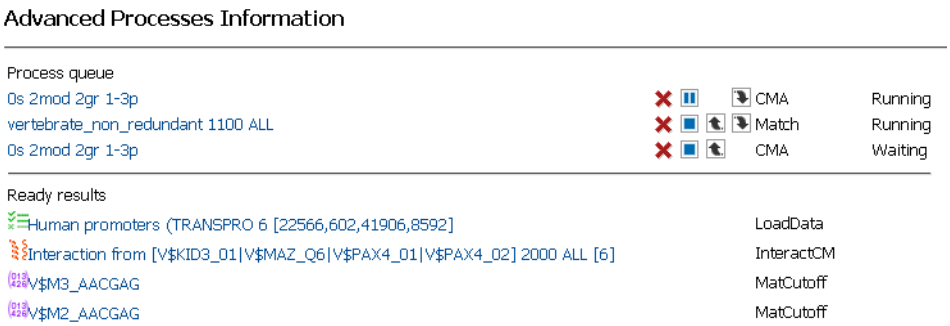
interface can also be used to move to the result of a newly completed analysis by clicking on its label when the monitor signals the results to be ready.

Figure 1.12 The process monitor with an active CMA process



You can take direct control over the process management through the [Advanced] button. In the control window, actions can be chosen for the process selected in the list. The button starts a waiting process possibly causing other running process to be paused. Conversely, any running process can be paused with the button. The button removes a process from the list. Finally, the and buttons on the right side of the list can be used to alter the order of processes in the queue. When a process replaces the process at the top of the list, the former top process will be paused.

Figure 1.13 The process control window



1.5 The workspace

The workspace displays data provided as input to any analysis tool as well as the analysis results. The contents of this frame depend on the active node in the project tree. Here we describe the general functionality of the workspace. Some additional options specific for certain tree nodes will be described in their corresponding sections.

1.5.1 Renaming project tree nodes

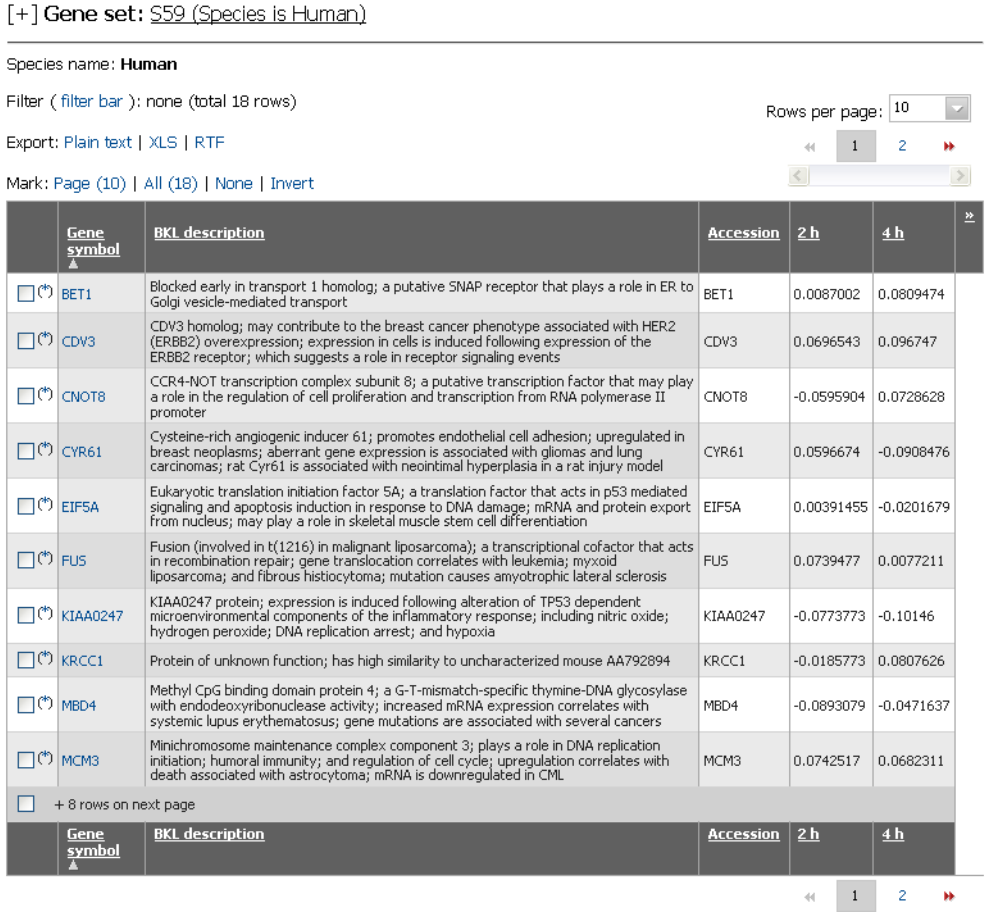
Each workspace provides a field to rename the active project tree node. To reach the editing field, select a node you want to rename in the project tree and click on its name on the top of the workspace. After editing the name in the line-editor and pressing the Rename button, ExPlain updates the interface with the new name.

1.5.2 Item origin information

By clicking on the [+] sign near the node name, you can see additional information. Four fields can be present here.

Parameters This field is shown for all analysis results. All parameters used by the algorithm are shown here. By pressing the [Run again] link or button new analyses with the same parameters can be launched.

Figure 1.14 Functions of the workspace



[+] Gene set: S59 (Species is Human)

Species name: **Human**

Filter (filter bar): none (total 18 rows)

Rows per page: 10

Export: Plain text | XLS | RTF

1 2

Mark: Page (10) | All (18) | None | Invert

	Gene symbol	BKL description	Accession	2 h	4 h
☐	BET1	Blocked early in transport 1 homolog; a putative SNAP receptor that plays a role in ER to Golgi vesicle-mediated transport	BET1	0.0087002	0.0809474
☐	CDV3	CDV3 homolog; may contribute to the breast cancer phenotype associated with HER2 (ERBB2) overexpression; expression in cells is induced following expression of the ERBB2 receptor; which suggests a role in receptor signaling events	CDV3	0.0696543	0.096747
☐	CNOT8	CCR4-NOT transcription complex subunit 8; a putative transcription factor that may play a role in the regulation of cell proliferation and transcription from RNA polymerase II promoter	CNOT8	-0.0595904	0.0728628
☐	CYR61	Cysteine-rich angiogenic inducer 61; promotes endothelial cell adhesion; upregulated in breast neoplasms; aberrant gene expression is associated with gliomas and lung carcinomas; rat Cyr61 is associated with neointimal hyperplasia in a rat injury model	CYR61	0.0596674	-0.0908476
☐	EIF5A	Eukaryotic translation initiation factor 5A; a translation factor that acts in p53 mediated signaling and apoptosis induction in response to DNA damage; mRNA and protein export from nucleus; may play a role in skeletal muscle stem cell differentiation	EIF5A	0.00391455	-0.0201679
☐	FUS	Fusion (involved in t(1216) in malignant liposarcoma); a transcriptional cofactor that acts in recombination repair; gene translocation correlates with leukemia; myxoid liposarcoma; and fibrous histiocytoma; mutation causes amyotrophic lateral sclerosis	FUS	0.0739477	0.0077211
☐	KIAA0247	KIAA0247 protein; expression is induced following alteration of TP53 dependent microenvironmental components of the inflammatory response; including nitric oxide; hydrogen peroxide; DNA replication arrest; and hypoxia	KIAA0247	-0.0773773	-0.10146
☐	KRCC1	Protein of unknown function; has high similarity to uncharacterized mouse AA792894	KRCC1	-0.0185773	0.0807626
☐	MBD4	Methyl CpG binding domain protein 4; a G-T-mismatch-specific thymine-DNA glycosylase with endodeoxyribonuclease activity; increased mRNA expression correlates with systemic lupus erythematosus; gene mutations are associated with several cancers	MBD4	-0.0893079	-0.0471637
☐	MCM3	Minichromosome maintenance complex component 3; plays a role in DNA replication initiation; humoral immunity; and regulation of cell cycle; upregulation correlates with death associated with astrocytoma; mRNA is downregulated in CML	MCM3	0.0742517	0.0682311
☐	+ 8 rows on next page				
	Gene symbol	BKL description	Accession	2 h	4 h

1 2

Figure 1.15 Click on the name to rename the node

[+] Gene set: (Fold change > 2)

Figure 1.16 Enter new name

[+] Gene set: (Fold change > 2) Rename Cancel

Origin This field contains a brief description of the item origin.

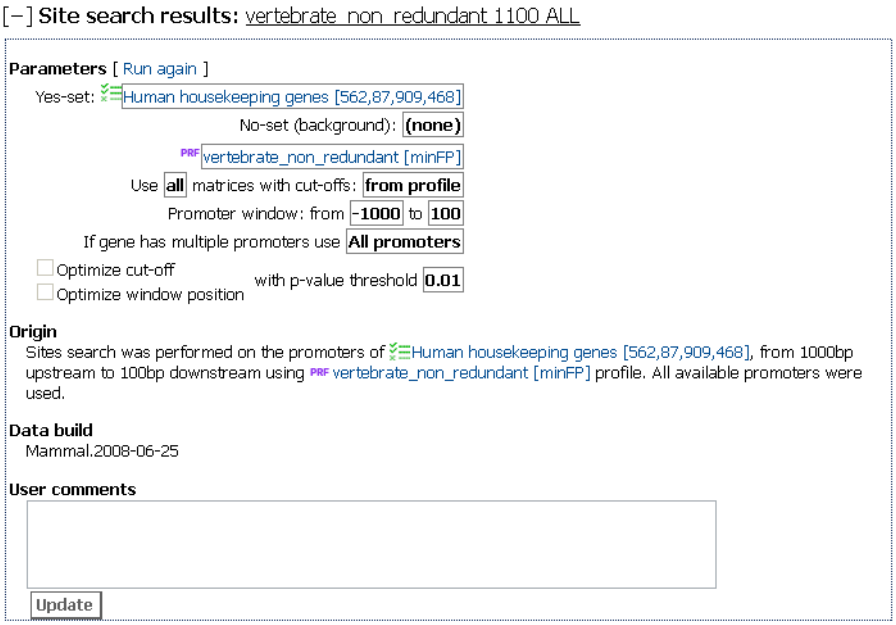
Data build This field shows the data file version used when this item was created. This can be useful when you update the data on an existing installation and want to know whether you performed some analysis using an older or newer version of the file.

User comments To store any notes in this field enter them in the comment window and press Update

1.5.3 Filtering table

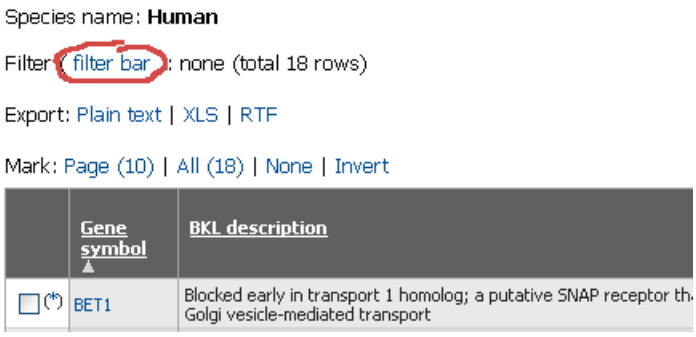
The Filter bar allows you to quickly filter any data table in ExPlain based on different conditions. Note that this feature only hides table rows that don't meet the condition. It doesn't change the data itself, and filtering can be removed easily.

Figure 1.17 Item origin information exemplified by MATCH results


[−] Site search results: vertebrate_non_redundant 1100 ALL
Parameters [Run again]
Yes-set: Human housekeeping genes [562,87,909,468]
No-set (background): (none)
vertebrate_non_redundant [minFP]
Use all matrices with cut-offs: from profile
Promoter window: from -1000 to 100
If gene has multiple promoters use All promoters
Optimize cut-off
Optimize window position
with p-value threshold 0.01
Origin
Sites search was performed on the promoters of Human housekeeping genes [562,87,909,468], from 1000bp upstream to 100bp downstream using vertebrate_non_redundant [minFP] profile. All available promoters were used.
Data build
Mammal.2008-06-25
User comments
Update

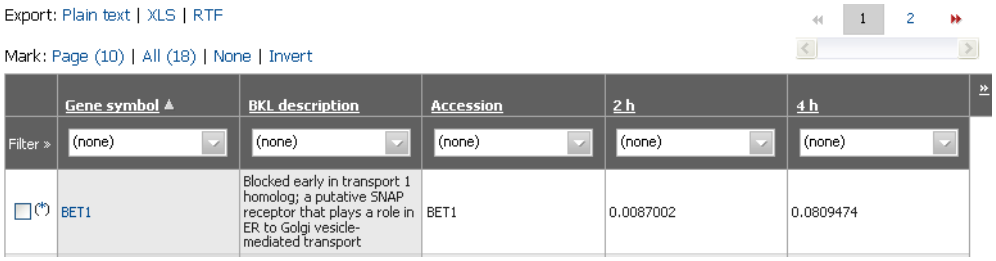
To turn on the filter bar, press the "filter bar" link above the table:

Figure 1.18


Species name: Human
Filter (filter bar): none (total 18 rows)
Export: Plain text | XLS | RTF
Mark: Page (10) | All (18) | None | Invert
Gene symbol
BKL description
BET1
Blocked early in transport 1 homolog; a putative SNAP receptor th
Golgi vesicle-mediated transport

Now the filter bar will appear below the column titles in the table.

Figure 1.19



When the filter is active, the filter bar is switched on by default. To hide it, press the "filter bar" link again.

To create a filter, click on the drop down menu and specify the filtering condition. For numerical data you can specify lower or upper limit, inner or outer range in the column. If you need to filter the column by an exact value, enter this value as a boundary for "in" range. For text data you can specify a substring to be found in the corresponding cells.

Hit "OK" button (or press Enter), and the filter will be applied. Now you will see the filtered table:

The filter row shows the actual filter and the number of rows which meet the filter conditions. If

Figure 1.20

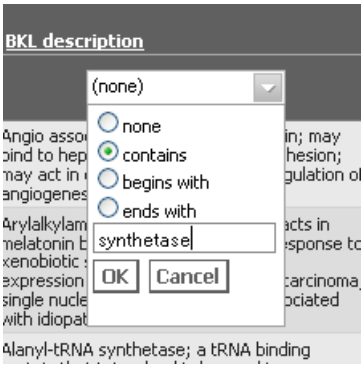
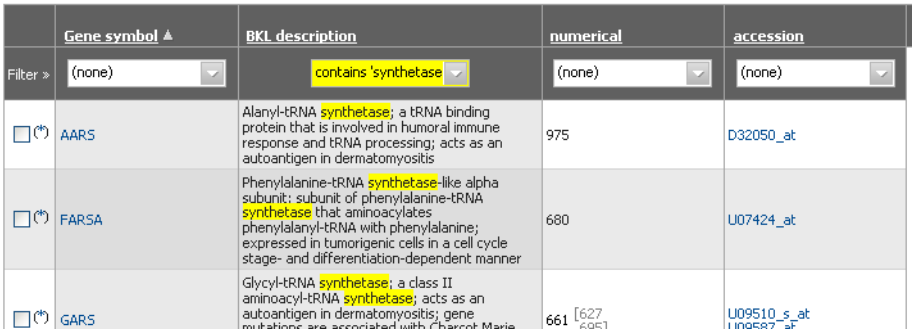


Figure 1.21


Gene symbol
BKL description
numerical
accession
Filter »
(none)
contains 'synthetase
(none)
(none)
AARS
Alanyl-tRNA synthetase; a tRNA binding protein that is involved in humoral immune response and tRNA processing; acts as an autoantigen in dermatomyositis
975
D32050_at
FARSA
Phenylalanine-tRNA synthetase-like alpha subunit; subunit of phenylalanine-tRNA synthetase that aminoacylates phenylalanyl-tRNA with phenylalanine; expressed in tumorigenic cells in a cell cycle stage- and differentiation-dependent manner
680
U07424_at
GARS
Glycyl-tRNA synthetase; a class II aminoacyl-tRNA synthetase; acts as an autoantigen in dermatomyositis; gene
661

filtering was performed on a text column, then the filtered substring is highlighted in each cell. You can add more conditions constraining other columns content. To clear the filter click the ”clear” link in the Filter row. To clear only one condition you can click the corresponding filter drop down menu, change condition to ”none” and press ”OK”.

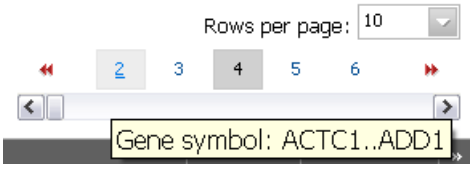
Filtering conditions persist when you select other tree items and after logout. When a filter is active, the ”Export” options will export only rows remaining after filtering, not the whole table. Also you may want to mark all rows and use the ”Get selected items” option in the menu as an alternative way to filter your data apart from ”Filter gene set by condition...”. Note that the filter bar works for any data set which is represented as a table, as opposed to ”Filter gene set by condition...” which works for gene sets only.

1.5.4 Adjusting the number of visible table rows and page navigation

The number of visible rows can be adjusted by selecting an appropriate value (10, 20, ..., 1000, all) from the dropdown box above the data table.

ExPlain will then immediately update the view accordingly. If the current view does not comprise all table rows, the interface provides a list of page links to navigate through the table. Hover the mouse pointer over the page number to see the corresponding value range of the sorting column.

Figure 1.22 Page navigation

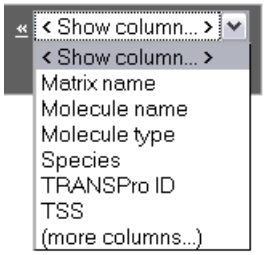


1.5.5 Sorting, renaming and showing/hiding table columns

Any table in the workspace can be sorted by the values of a column by clicking on the corresponding column title.

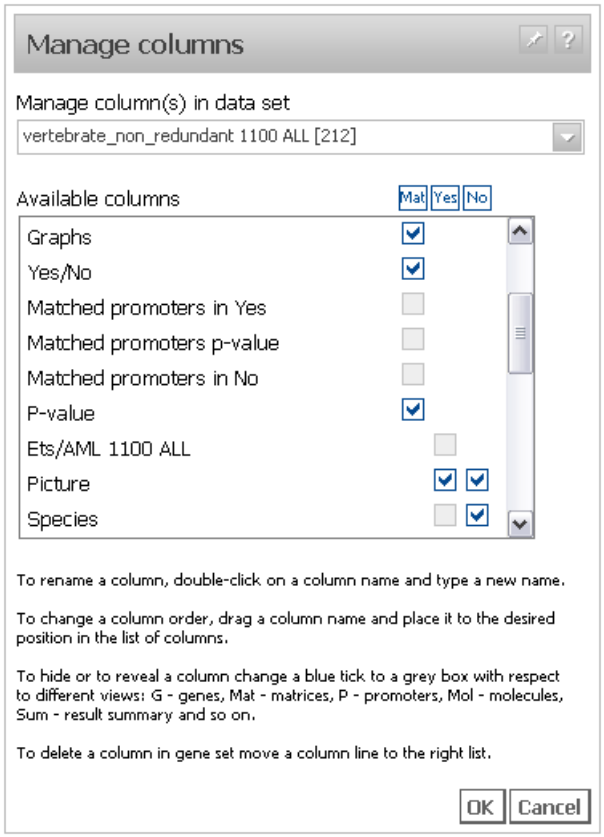
If the table is sorted, an arrow next to the title indicates whether the items are arranged in ascending or descending order. Clicking the title of a sorted column switches to the opposite order. Furthermore, any column can be hidden by clicking on the icon that appears next to the column title when placing the mouse over it. Hidden columns can be reintroduced to the table by clicking on their names in the list on the right side of the header row.

Figure 1.23 Hidden column selector



An advanced mechanism for hiding, removing, renaming and sorting columns is available trough the ”Manage columns in gene set” menu link in the ”Data” menu. The column management dialog is displayed on the figure below.

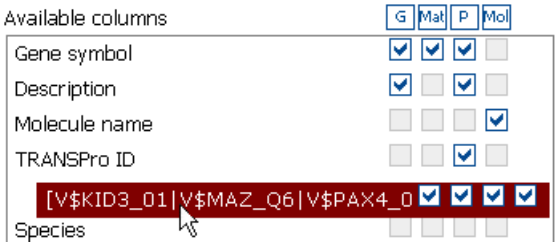
Figure 1.24 Advanced column management


Manage columns
Manage column(s) in data set
vertebrate_non_redundant 1100 ALL [212]
Available columns
Mat
Yes
No
Graphs
Yes/No
Matched promoters in Yes
Matched promoters p-value
Matched promoters in No
P-value
Ets/AML 1100 ALL
Picture
Species
To rename a column, double-click on a column name and type a new name.
To change a column order, drag a column name and place it to the desired position in the list of columns.
To hide or to reveal a column change a blue tick to a grey box with respect to different views: G - genes, Mat - matrices, P - promoters, Mol - molecules, Sum - result summary and so on.
To delete a column in gene set move a column line to the right list.
OK
Cancel

First select an item to manage, note that not all item types allow column management. All available columns will be listed in the column control. Different views are designated in the top line. If you point your mouse pointer over a button, a hint with a view name will appear. Marked check boxes mean that a column is visible in the corresponding view. You can easily change the visibility by selecting or deselecting check boxes. To change the column order just select a column and drag-and-drop it to a desirable place in the column list as shown in the figure. To rename a column click twice on it’s name

and type a new name inthe text box. To remove a column from view select it and drag it to the right column container (not available for all item types). To cancel this action drag the column back.

Figure 1.25



1.5.6 Customization of the column content display

The way your data is represented in the workspace can be altered individually for each column by pressing on the icon, which appears when hovering your mouse pointer over the column title. Or, for several columns, via the "Rename/customize columns" dialog from the "Data" menu. Figure 1.26 shows options for the four column types available. Only one type will be displayed for each column.

Number format Here you can select normal (e.g. 0.0015) or scientific (e.g. 1.5e-3) number format, number of decimal digits, and rules for the effective value calculation. The effective value is that which will be used for the analyses and sorting when certain entity has several corresponding annotations. For example, when several Affymetrix identifiers with different expression values are matched to a single gene, you can select to use maximum, minimum, average, sum or extreme of all values as an effective expression value for this gene.

Text format With the "Merge duplicate strings" option selected, only unique text strings will be shown. Each string will be shown at a new line.

DB-link format Here you can change the number of identifiers shown on the screen, separation type and presence or absence of duplicate entries.

Categorical format Within this dialog, categories can be freely assigned numbers in order to range them. To modify an assignment, choose a category from the list with all currently assigned category-number pairs in the selected gene set, and enter the new value in the editor field. You can also adjust effective value rules here.

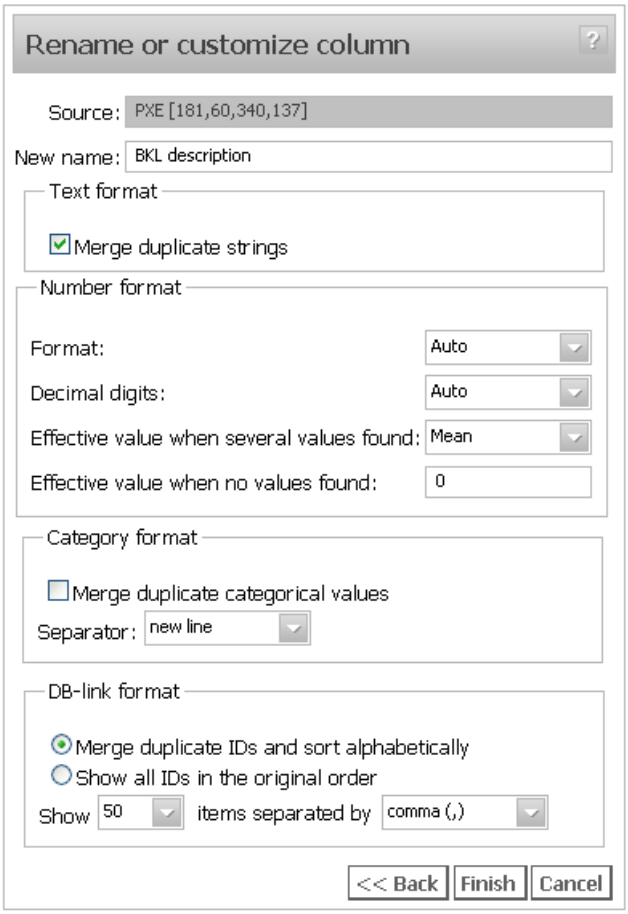
1.5.7 Selecting rows and further actions

Table presentations of ExPlain usually contain a checkbox column that can be used to select data rows for different actions, such as extraction of subsets. Actions available for the selected rows (for example "Get selected genes as a gene set" for the Gene set or "Show site maps for selected matrices" for the MATCH result) can be found in the custom menu, specific for each tree node, as well as on the toolbar. More information is given in the corresponding documentation section.

Additional checkboxes appear when the data table size exceeds the number of rows per page. All data rows on previous pages can be selected by a single checkbox in the top row. Data rows on following pages are available through a checkbox in the bottom row.

As a convenient function, checkbox columns provide range selection and deselection options, so that the status of the last modified checkbox can be transferred to a whole range of rows by pressing the [Shift] key while selecting the top or bottom row of the desired range. Marked data rows can then be extracted to a new subset node by pressing the appropriate menu link.

Figure 1.26 Customization of the column content display



Marking options right above the table provide additional possibilities. You can select all rows on the current page, or all rows in the table if it has more than one page, clear all selections, and invert selections by turning selected rows to deselected and vice versa.

If an action is applied with no rows selected, all rows from the table will be used as an input.

1.5.8 Data export

The workspace provides you with the opportunity to export your data in different formats, such as tab-delimited text or a Microsoft™ Excel spreadsheet, for use outside of ExPlain. Certain tree nodes have additional export formats. For example "Genome intervals" can be exported as a BED file.

Figure 1.27 Data export links

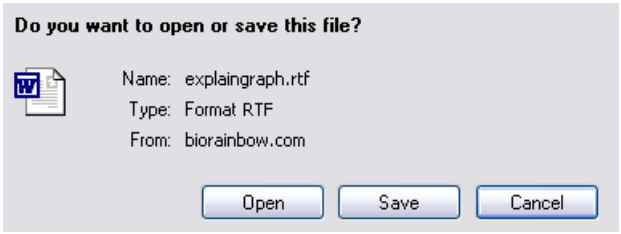
Export: Plain text | XLS | RTF

Some ExPlain plots can be exported in RTF format as a vector graphics image, so that it can be resized with other applications such as Microsoft™ Word. Press the "RTF" link in Export: RTF | XLS option above the plot and save the file or open it using the "File download" dialog. The "XLS" link exports the plot to Microsoft™ Excel stylesheets as a set of values.

1.5.9 Interface preferences

The "Preferences..." menu option of the "View" menu launches the adjustment dialog shown on Figure 1.29. You can change the way the tree is displayed (more or less space between tree entries), the item statistics and date of creation. The date format and time zone can also be specified.

Figure 1.28 Data export links

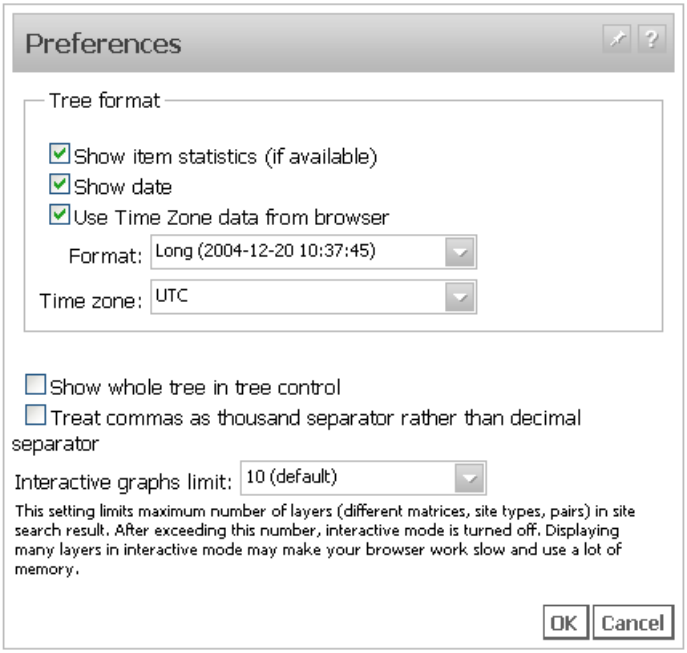


If the "Show whole tree in tree control" option is chosen, the tree-node selection control (Figure 1.8) will display all tree items, otherwise only items of acceptable type will be shown. The "Interactive graphs limit" option limits the maximum number of layers (different matrices, site types, pairs) in site search results in the matrix legend (see Section 5.2.1). If this number is exceeded, interactive mode is turned off.

WARNING: Displaying many layers in interactive mode may make your browser work slow and use a lot of memory.

You can also set commas as thousand separators rather than decimal separators.

Figure 1.29 Preferences dialog


Preferences
Tree format
Show item statistics (if available)
Show date
Use Time Zone data from browser
Format: Long (2004-12-20 10:37:45)
Time zone: UTC
Show whole tree in tree control
Treat commas as thousand separator rather than decimal separator
Interactive graphs limit: 10 (default)
This setting limits maximum number of layers (different matrices, site types, pairs) in site search result. After exceeding this number, interactive mode is turned off. Displaying many layers in interactive mode may make your browser work slow and use a lot of memory.
OK
Cancel

Chapter 2

Analytical workflows and Wizard mode

2.1 Workflow

The Wizard mode provides you with a set of analytical workflows that will allow you to:

Perform comprehensive promoter analyses.

Predict prospective drug targets.

Reveal disease biomarkers.

Explain mechanisms of drug toxicity.

To enter the wizard mode you can choose one of the following options:

- Press the toolbar icon.

- Click the ”Workflow” menu link in the ”View” menu.

- Click on the ”Switch to workflow mode” link on the start page.

ThefFirst part of the workflow will upload new data. You can choose to load your data in several common formats, as described in Section 2.2. After the data import is scheduled you will be forwarded to the workflow screen where newly loaded data will be preselected for the analyses. If you already have uploaded data on which you want to run one of the workflows, you canskip the data loading and proceed directly to the workflow section (Section 2.3).

2.2 Data loading

This chapter describes various data formats recognized by ExPlain via the Wizard.

Choose the most siutable data structure based on the short descriptions present on the workflow icons. Clicking on any icon will expand it to show the list of specific data formats supported. Click on one of the links to upload a data file or copy/paste data manualy.

2.2.1 Gene set

If you have an already precompiled set of co-regulated genes or sequences you can upload it using one of the ”gene set” options. We recommend you to use a set of no more that 1000 genes, the optimal size would be between 100 and 200 genes.

Figure 2.1 Data loading page

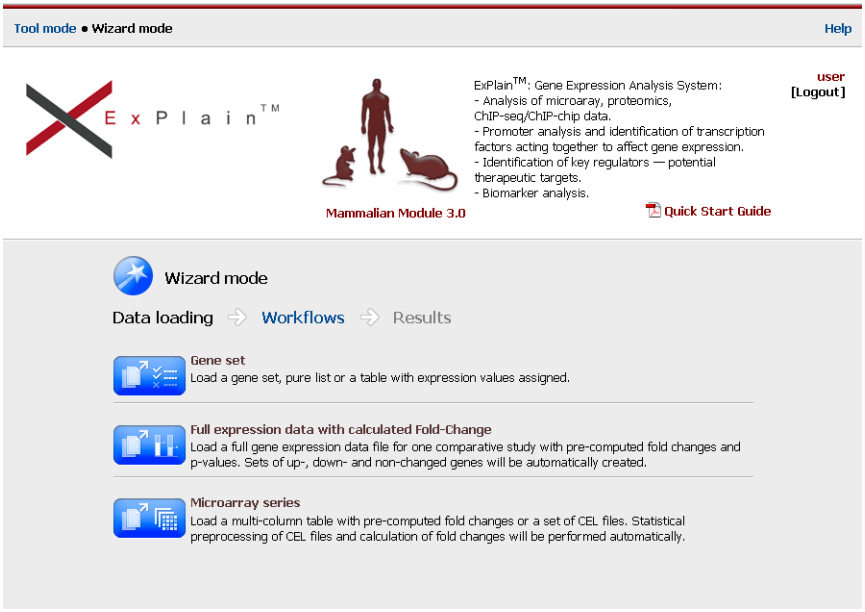

Tool mode • Wizard mode
Help
E x P l a i n™
ExPlain™: Gene Expression Analysis System:
- Analysis of microarray, proteomics,
ChIP-seq/ChIP-chip data.
- Promoter analysis and identification of transcription
factors acting together to affect gene expression.
- Identification of key regulators — potential
therapeutic targets.
- Biomarker analysis.
user
[Logout]
Mammalian Module 3.0
Quick Start Guide
Wizard mode
Data loading Workflows Results
Gene set
Load a gene set, pure list or a table with expression values assigned.
Full expression data with calculated Fold-Change
Load a full gene expression data file for one comparative study with pre-computed fold changes and p-values. Sets of up-, down- and non-changed genes will be automatically created.
Microarray series
Load a multi-column table with pre-computed fold changes or a set of CEL files. Statistical preprocessing of CEL files and calculation of fold changes will be performed automatically.

Figure 2.2 Data loading page

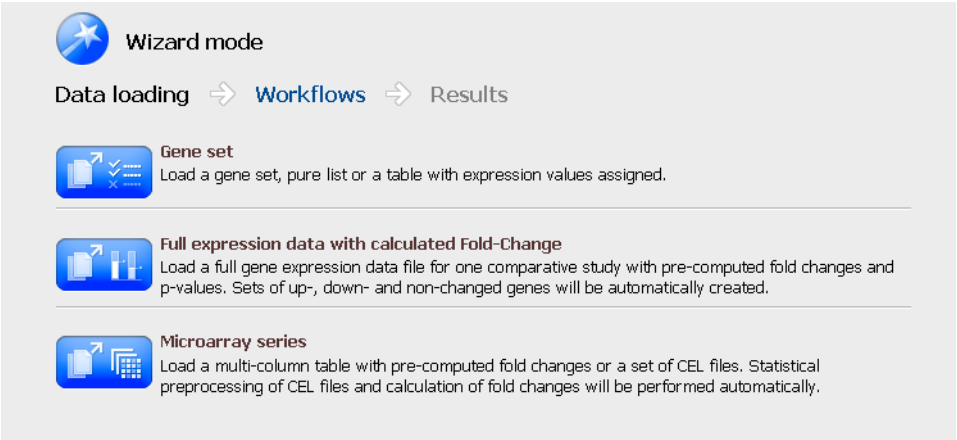
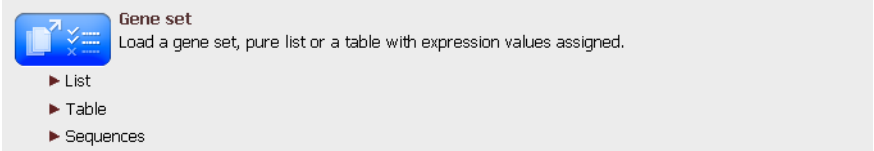

Wizard mode
Data loading Workflows Results
Gene set
Load a gene set, pure list or a table with expression values assigned.
Full expression data with calculated Fold-Change
Load a full gene expression data file for one comparative study with pre-computed fold changes and p-values. Sets of up-, down- and non-changed genes will be automatically created.
Microarray series
Load a multi-column table with pre-computed fold changes or a set of CEL files. Statistical preprocessing of CEL files and calculation of fold changes will be performed automatically.

Figure 2.3 Gene set formats



List: copy and paste or type a list of identifiers separated by space, tab, comma, end of line, etc. as shown on the figure below.

Table:load a tab-separated text format file, or Exel worksheet, that contains a single column with a list of gene identifiers, one identifier per row.

When the table is uploaded you will see it on the screen. You can change column names and types and apply filters to shorten the list. Make sure that the column containing the identifiers is recognized correctly as "accession" and press >> Finish importing.

Sequences: load a list of sequences in FASTA or EMBL format.

Figure 2.4

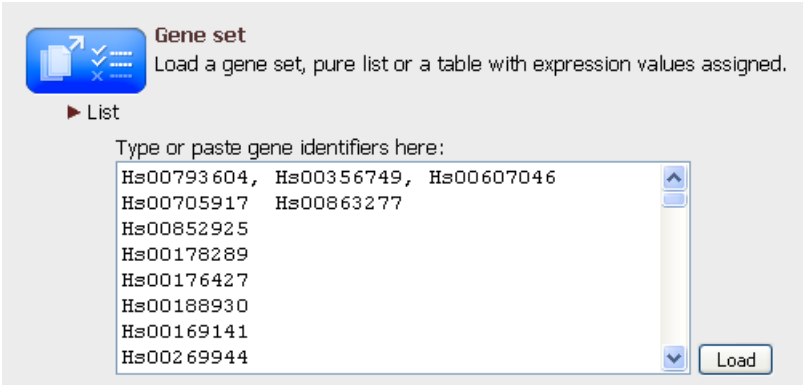
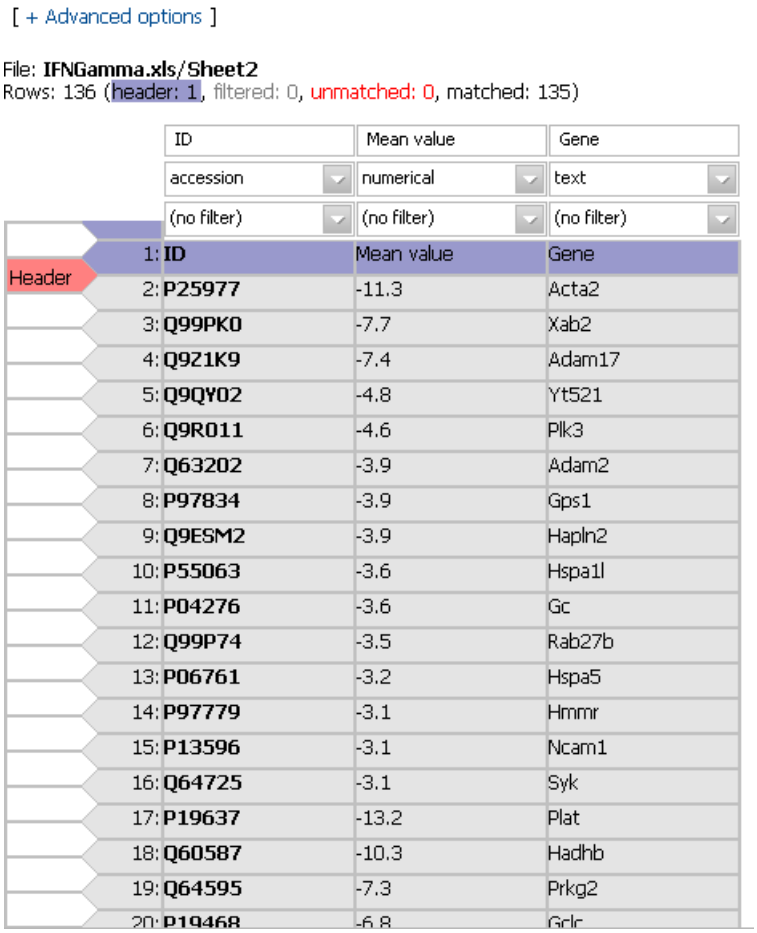


Figure 2.5



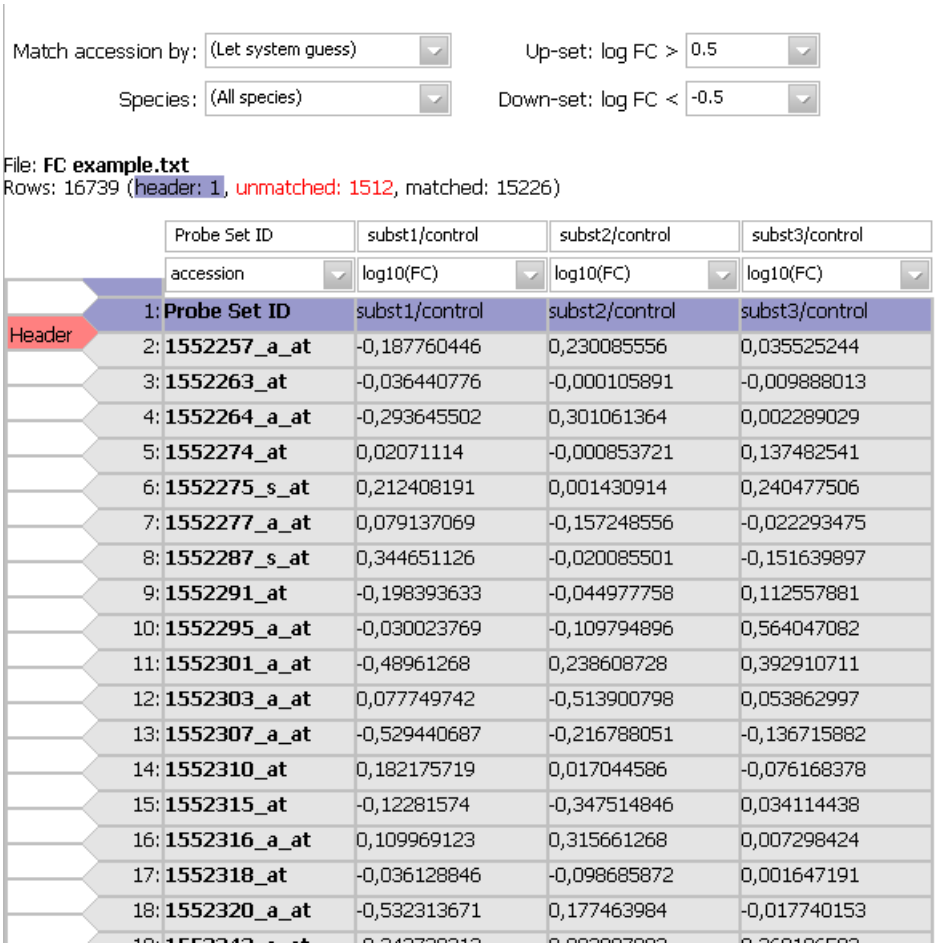
2.2.2 List with fold change

Use this option when you have a file with gene identifiers and precalculated fold change (and p-value, optionally) columns, in either tab-separated text or excel format. If you upload a list of this kind, ExPlain will display information on the number of matched rows and sizes of the Up- and Down- regulated gene sets which will be created and used in the analyses. Adjust the filter options to have a reasonable amount of genes in these sets (between 20 and 500) and press >> Finish importing.

2.2.3 Microarray series

Table: Use this option when you have a plain text file or an Excel table containing a list of genes with several fold change columns assigned. For each fold change column, 2 sets, Up- and Down-regulated genes will be created.

Figure 2.6 A data set with several pre-calculated Fold Change columns



CEL files: Here you can load CEL files from one experiment archived in ZIP, TAR, or Gzip TAR format. After the file is uploaded you will see a dialog that allows you to define groups of arrays for further comparison, as shown in Figure 2.7. When the name of the comparison (factor) is defined, you will need to assign arrays to experiment or control groups. Later on in the workflow ExPlain will analyze genes differentially expressed between this 2 groups.

If you want to analyse several comparisons (for example "treatment 1 vs. control", "teatment 2 vs. control" and "treatment 1 vs. treatment 2") create several factors, marking for each certain arrays as experiment or control.

When the assignment is done press >> Next.

You will then see the parameters for the statistical processing of the CEL files. Here you can in most cases leave the default values as they are. FC filtering options indicate the threshold that will be used to define up and down regulated sets.

2.3 Workflow mode

The full upstream analysis (common regulators in signalling network) workflow will schedule a number of processes to find over represented functional groups in your gene sets, predict relevant transcription factors, upstream regulators and affected pathways.

First you have to select analyzed (Yes) and background (No) sets. If you went through a data loading step in the workflow, the results of the previous step will be already preselected. In the form below

Figure 2.7 Factor-level assignment in workflow

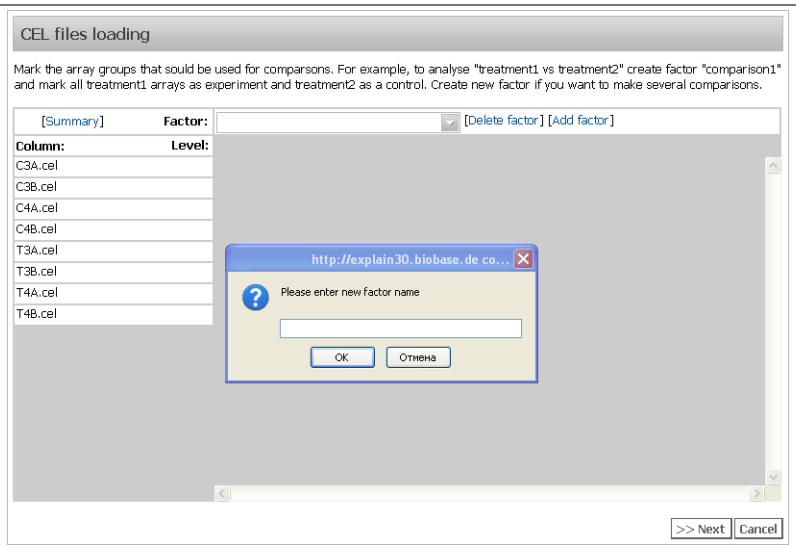


Figure 2.8 Factor-level assignment in workflow

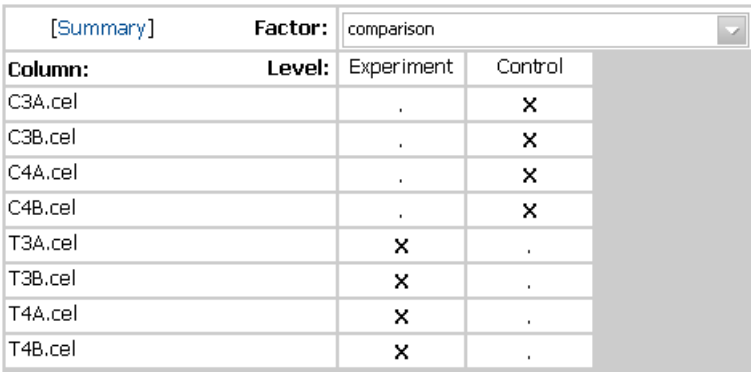
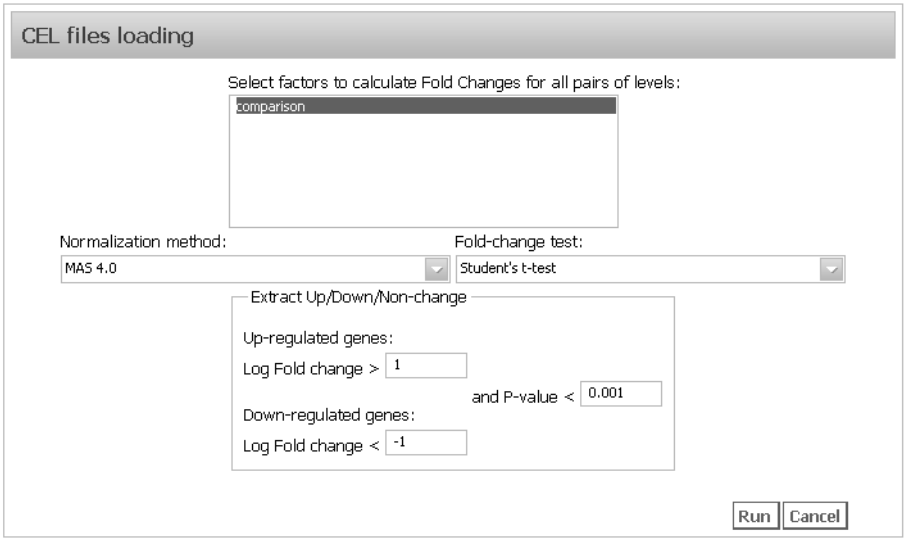


Figure 2.9 Statistical analyses parameters window



default parameters are selected. You can change the minimal number of genes to consider for the functional classification, and the size of the promoter sequence to analyse

After pressing the Run button you will be redirected to the report page, which at the beginning will

Figure 2.10 Full upstream analysis link

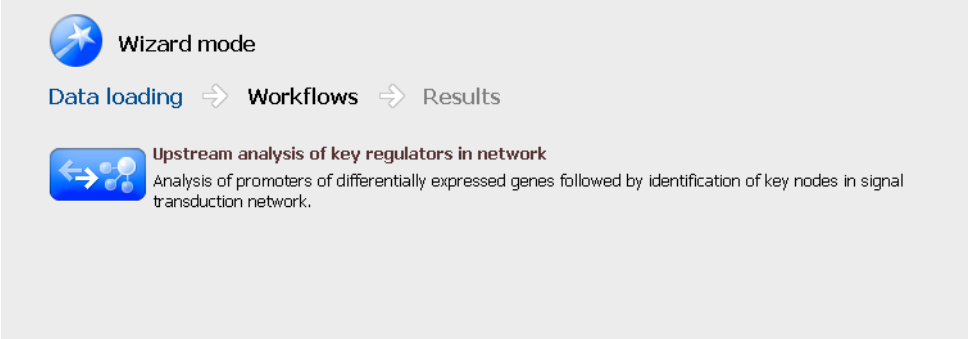
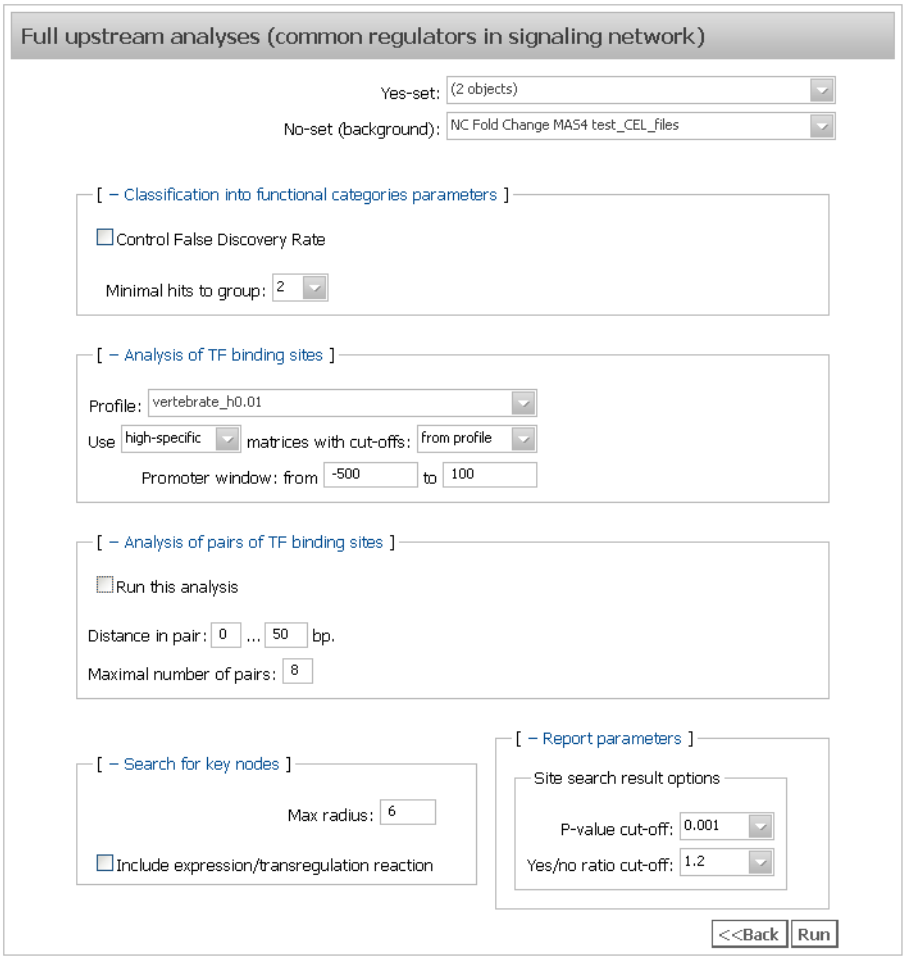

Wizard mode
Data loading
Workflows
Results
Upstream analysis of key regulators in network
Analysis of promoters of differentially expressed genes followed by identification of key nodes in signal transduction network.

Figure 2.11 Full upstream analysis parameters dialog


Full upstream analyses (common regulators in signaling network)
Yes-set: (2 objects)
No-set (background): NC Fold Change MAS4 test_CEL_files
[– Classification into functional categories parameters]
Control False Discovery Rate
Minimal hits to group: 2
[– Analysis of TF binding sites]
Profile: vertebrate_h0.01
Use high-specific matrices with cut-offs: from profile
Promoter window: from -500 to 100
[– Analysis of pairs of TF binding sites]
Run this analysis
Distance in pair: 0 ... 50 bp.
Maximal number of pairs: 8
[– Search for key nodes]
Max radius: 6
Include expression/transregulation reaction
[– Report parameters]
Site search result options
P-value cut-off: 0.001
Yes/no ratio cut-off: 1.2
<<Back
Run

display only a note on the number of scheduled processes and the progressbar of the currently running one. In the process tree you will see new nodes for all the scheduled processes.

Each data set will get a tree branch under it which will include Transcription Factor (TF) site prediction (f-match) results, a selection of relevant TFs, and the result of the upstream search in relevant signal transduction networks to find key molecules, as well as network visualization nodes. The best key molecules will be then extracted, combined with the main data set and passed to the pathway classification analysis to find relevant pathways. The figures below show one of the branches in waiting state (Figure 2.14) and ready state (Figure 2.15).

When you have several data sets in the analysis, to facilitate the comparison of the results ExPlain will make a table with a summary of all analysis of the same type.

Figure 2.12

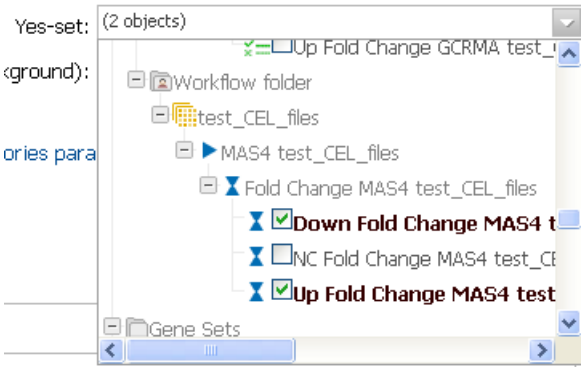


Figure 2.13

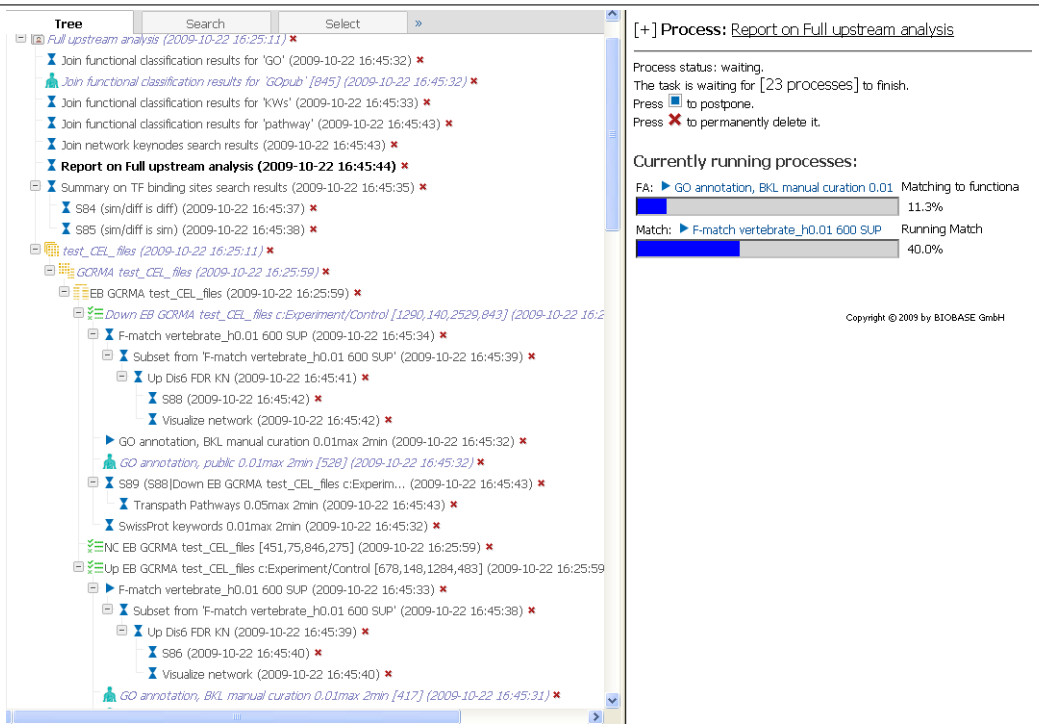

Tree
Search
Select
Full upstream analysis (2009-10-22 16:25:11)
Join functional classification results for 'GO' (2009-10-22 16:45:32)
Join functional classification results for 'GOpub' [845] (2009-10-22 16:45:32)
Join functional classification results for 'KWs' (2009-10-22 16:45:33)
Join functional classification results for 'pathway' (2009-10-22 16:45:43)
Join network keynodes search results (2009-10-22 16:45:43)
Report on Full upstream analysis (2009-10-22 16:45:44)
Summary on TF binding sites search results (2009-10-22 16:45:35)
S84 (sim/diff is diff) (2009-10-22 16:45:37)
S85 (sim/diff is sim) (2009-10-22 16:45:38)
test_CEL_files (2009-10-22 16:25:11)
GCRMA test_CEL_files (2009-10-22 16:25:59)
EB GCRMA test_CEL_files (2009-10-22 16:25:59)
Down EB GCRMA test_CEL_files c:Experiment/Control [1290,140,2529,843] (2009-10-22 16:2
F-match vertebrate_h0.01 600 SUP (2009-10-22 16:45:34)
Subset from 'F-match vertebrate_h0.01 600 SUP' (2009-10-22 16:45:39)
Up Dis6 FDR KN (2009-10-22 16:45:41)
S88 (2009-10-22 16:45:42)
Visualize network (2009-10-22 16:45:42)
GO annotation, BKL manual curation 0.01max 2min (2009-10-22 16:45:32)
GO annotation, public 0.01max 2min [528] (2009-10-22 16:45:32)
S89 (S88|Down EB GCRMA test_CEL_files c:Experim... (2009-10-22 16:45:43)
Transpath Pathways 0.05max 2min (2009-10-22 16:45:43)
SwissProt keywords 0.01max 2min (2009-10-22 16:45:32)
NC EB GCRMA test_CEL_files [451,75,846,275] (2009-10-22 16:25:59)
Up EB GCRMA test_CEL_files c:Experiment/Control [678,148,1284,483] (2009-10-22 16:25:59
F-match vertebrate_h0.01 600 SUP (2009-10-22 16:45:33)
Subset from 'F-match vertebrate_h0.01 600 SUP' (2009-10-22 16:45:38)
Up Dis6 FDR KN (2009-10-22 16:45:39)
S86 (2009-10-22 16:45:40)
Visualize network (2009-10-22 16:45:40)
GO annotation, BKL manual curation 0.01max 2min [417] (2009-10-22 16:45:31)
[+] Process: Report on Full upstream analysis
Process status: waiting.
The task is waiting for [23 processes] to finish.
Press to postpone.
Press to permanently delete it.
Currently running processes:
FA: GO annotation, BKL manual curation 0.01
Matching to functiona
11.3%
Match: F-match vertebrate_h0.01 600 SUP
Running Match
40.0%
Copyright © 2009 by BIOBASE GmbH

Figure 2.14

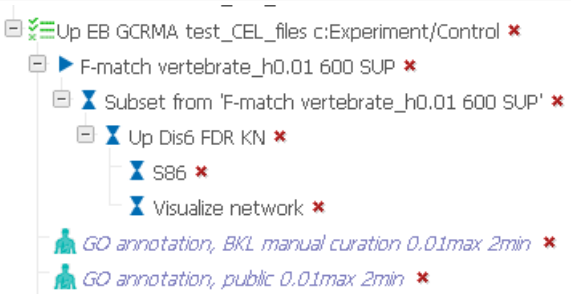

Up EB GCRMA test_CEL_files c:Experiment/Control
F-match vertebrate_h0.01 600 SUP
Subset from 'F-match vertebrate_h0.01 600 SUP'
Up Dis6 FDR KN
S86
Visualize network
GO annotation, BKL manual curation 0.01max 2min
GO annotation, public 0.01max 2min

Figure 2.15

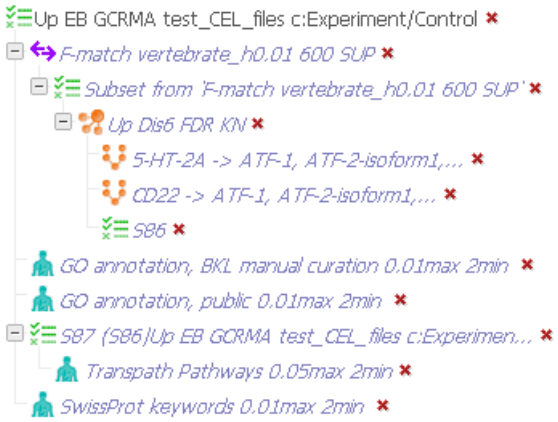

Up EB GCRMA test_CEL_files c:Experiment/Control
F-match vertebrate_h0.01 600 SUP
Subset from 'F-match vertebrate_h0.01 600 SUP'
Up Dis6 FDR KN
5-HT-2A -> ATF-1, ATF-2-isoform1,...
CD22 -> ATF-1, ATF-2-isoform1,...
S86
GO annotation, BKL manual curation 0.01max 2min
GO annotation, public 0.01max 2min
S87 (S86)Up EB GCRMA test_CEL_files c:Experimen...
Transpath Pathways 0.05max 2min
SwissProt keywords 0.01max 2min

Figure 2.16 Summary of several functional classifications

Full upstream analysis
Join functional classification results for 'GO'
Join functional classification results for 'GOpub'
Join functional classification results for 'KWs'
Join functional classification results for 'pathway'
Join network keynodes search results
Summary on TF binding sites search results

Chapter 3

Gene sets

This chapter explains how to load gene set data into ExPlain and describes ways to manipulate data sets and data representation details.

3.1 Loading data into the ExPlain system

The first step to analyse any data set with ExPlain is to upload it to the ExPlain system. There are several ways to create a data set in ExPlain. A user can:

load data as a file,

create a set of data by pasting a list of accession numbers into a text dialog window,

create a subset from results of other analysis previously created with ExPlain, or

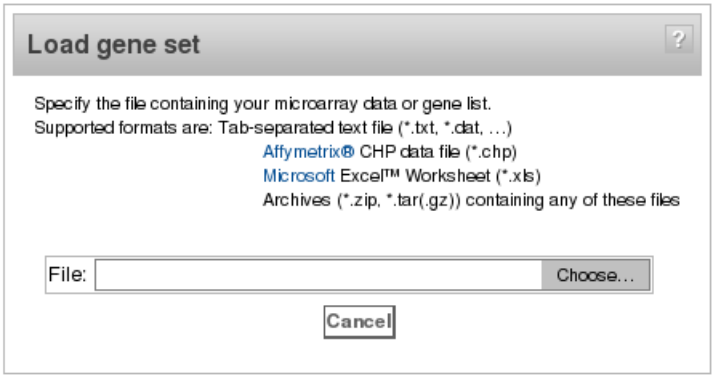
import search results from BKL.

3.1.1 Load gene set dialog

The process of data loading is guided step-by-step, assuring that data are represented in an easy to understand relational schema.

The first window of the "Load gene set" dialog is shown in Figure 3.1. To load any set of data from a file, specify the file name. You will then automatically be directed to the detailed view of the data, where you can control and change parameters recognized by the system.

Figure 3.1 "Load geneset" dialog


Load gene set
Specify the file containing your microarray data or gene list.
Supported formats are: Tab-separated text file (*.txt, *.dat, ...)
Affymetrix® CHP data file (*.chp)
Microsoft Excel™ Worksheet (*.xls)
Archives (*.zip, *.tar(.gz)) containing any of these files
File:
Choose...
Cancel

3.1.1.1 Supported species

At the moment the mamalian module of ExPlain supports data from human, rat, and mouse. ExPlain-Plant accepts data from Arabidopsis thaliana, Oryza sativa, and Glycine max.

3.1.1.2 Supported file formats

ExPlain supports four data formats - Affymetrix™ <http://www.affymetrix.com> CHP files, Affymetrix Exon genecore CHP files, ASCII text files, and Microsoft™ <http://www.microsoft.com> Excel™ files. Any of these files can also be loaded as an archive (*.zip or *.tar.gz). Within text and Excel™ files, data are assumed to be presented as a table with one entity per row. An arbitrary range of non-data header rows, e.g. containing notes or a column title bar, may be present at the top of the table. In ASCII text files, data columns should be delimited by a single tabulator, where each row should contain the same number of tabulations, so that columns with undefined values are never skipped in individual rows.

NOTE

Currently ExPlain does not support Unicode <http://www.unicode.org> encoding in Excel™ files affecting, e.g., Asian or Cyrillic characters. Such files should be saved with the Western encoding. Otherwise proper processing of the data can not be guaranteed. Furthermore files from Microsoft™ <http://www.microsoft.com> Excel™ 2007 are presently not supported by the system.

3.1.2 Import options of geneset loading dialog

The import options dialog window displays the first rows of the loaded data. In case you did not select the correct file, or wish to load another file, you can return to the previous step by pressing the << Load another file button. When the source file name is correct, you can select the name of the tree folder where your data will be loaded in the "Destination" drop-down menu.

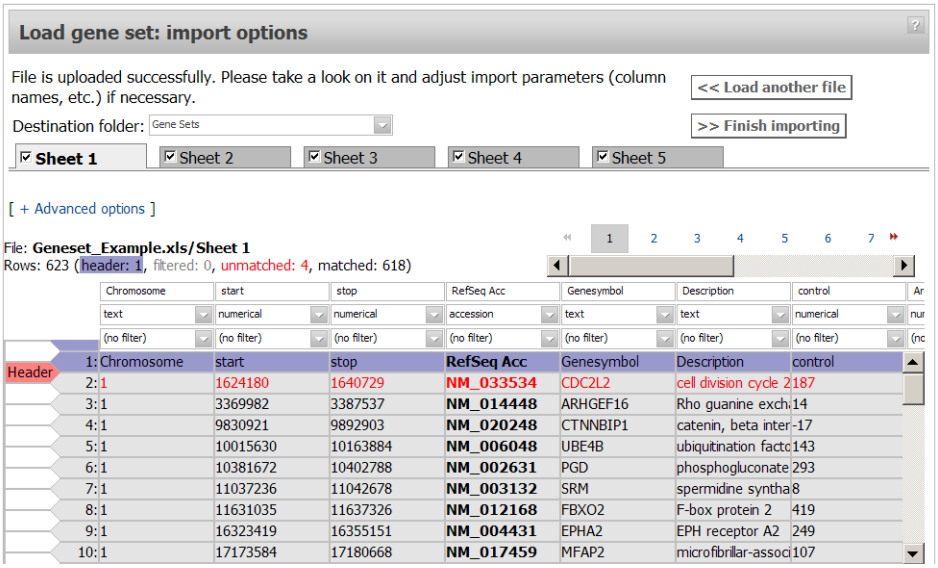
If you are loading an Excel™ file that contains several worksheets, each sheet will be represented by a sheet button, e.g. ☑ Sheet 2, in the upper part of the "Import options" dialog (Figure 3.2). The buttons are labeled with the sheet names and contain check boxes. To exclude worksheets from the data loading process, deselect the corresponding check boxes. If more than one worksheet is selected, they will be processed one after another (empty sheets will be skipped).

The dialog window also displays the head of your data table and asks you to specify the header/data separator. ExPlain guesses the location of header rows, such as a column title bar. If its suggestion is not correct, the header range can be adapted by moving the mouse pointer over the appropriate row separator of the data table and clicking on it. Rows to be excluded as part of the header are distinguished from the data section by a blue background.

TIP

Click on the separator above the first row if your file contains no header rows.

Figure 3.2 Import options dialog window


Load gene set: import options
File is uploaded successfully. Please take a look on it and adjust import parameters (column names, etc.) if necessary.
<< Load another file
Destination folder: Gene Sets
>> Finish importing
Sheet 1
Sheet 2
Sheet 3
Sheet 4
Sheet 5
[+ Advanced options]
File: Geneset_Example.xls/Sheet 1
Rows: 623 (header: 1, filtered: 0, unmatched: 4, matched: 618)
Header
1: Chromosome start stop RefSeq Acc Genesymbol Description control
2: 1 1624180 1640729 NM_033534 CDC2L2 cell division cycle 2 187
3: 1 3369982 3387537 NM_014448 ARHGEF16 Rho guanine exch 14
4: 1 9830921 9892903 NM_020248 CTNNBIP1 catenin, beta inter -17
5: 1 10015630 10163884 NM_006048 UBE4B ubiquitination facto 143
6: 1 10381672 10402788 NM_002631 PGD phosphogluconate 293
7: 1 11037236 11042678 NM_003132 SRM spermidine syntha 8
8: 1 11631035 11637326 NM_012168 FBXO2 F-box protein 2 419
9: 1 16323419 16355151 NM_004431 EPHA2 EPH receptor A2 249
10: 1 17173584 17180668 NM_017459 MFAP2 microfibrillar-associ 107

NOTE

The first line above the data section is not completely discarded, but is used to specify default column titles, which can then be customized in the next step.

NOTE

If the header section exceeds the range of visible rows and, therefore, the start of the data section can not be specified in the initial presentation, click on the number ”2” in the list of numbers above the data table to view the next entries.

The final step involves categorization of the data columns and, if necessary, editing of column titles. Each column can be assigned to one of the following eight categories:

Accession An accession column contains database accession numbers uniquely identifying the sequences corresponding to the data rows. Accession numbers can be derived from Affymetrix™, Entrez Gene, Unigene, any BIOBASE database or others. A complete list of available sequence sources is in Table 3.1.

NOTE

Only one column may be assigned to this primary accession category, while several columns can be assigned to other categories. In case you want to define an additional accession column use the category ”additional accession”.

Additional accession This category allows you to define a column of the data set as an additional accession column, that besides the primary accession column (required for all data sets) servers as a secondary accession set.

Categorical A categorical column classifies expression values into a discrete set of categories. Typically, such categories are integers or strings of one or two letters, e.g. the set of categories {I,MI,NC,MD,D}.

Numerical A numerical column contains real-numbered expression data such as raw intensities, normalized intensities or expression change ratios.

Text Text columns contain all sorts of text annotation.

Db-link The Db-link column contains various database accession numbers. They are recognized by the system and are provided with the links to the corresponding database entries.

Location This category marks columns with the chromosomal location information to provide proper sorting of data by this column (like Chr.1 4534456 or Chr:1p21).

Exclude This category marks a column to be skipped in the ExPlain import process.

Table 3.1 provides a list of accession numbers sources, the required format, and an example.

Table 3.1 Supported databases

Database	Format	Example
Affymetrix™	Probe set ID	41657_at
EMBL[13]/GenBank[14]/DDBJ[15]	Accession number	U15637
Ensembl[16]	Gene ID	ENSG00000118046
Entrez Gene[17]	Gene ID	5966
Entrez Protein[18]	Accession number	AAC03365
HGNC	Approved gene symbol	STK11
IPI[19]	Accession number	IPI00029773
MGI	Approved gene symbol	Arnt
RefSeq	Accession number	NM_002908
RGNC	Approved gene symbol	Arha2
TRANSFAC	Accession number	T00168
TRANSPATH	Accession number	MO000019368
TRANSPro	Accession number	MMU_81230
UniGene[20]	Cluster ID	Hs.279920
UniProt[21]	Accession number	Q04864
Interpro	Other ID	IPR000001
SwissProt/TrEmbl	Other ID	1433B_HUMAN
DBTSS	Other ID	29R00318
Fantom	Other ID	T01F000E467B

The categories can be selected from boxes above each column Figure 3.2. Column titles can be edited in line-editors above the category boxes.

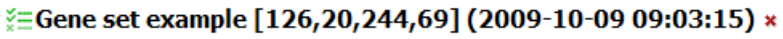
After pressing the [+ Advanced options] link you can set several additional options. You can specify the name of the database and species for matching accession numbers. If you made some changes, and want to save them for the future use, fill in a name for the preset and click on the Save preset button. It is also possible to make use of a preset saved before, by selecting it in the "Load Preset" drop down menu. When an Excel™ file with several worksheets is loaded, the specified settings can be applied to all other worksheets by clicking on the Copy settings to other sheets button. After pressing the >> Finish importing button, you are redirected to the ExPlain interface and your data is uploaded to the system.

NOTE

Due to pre-processing for further analyses, data transfer may take some time. This is also influenced by the size of the data set or other jobs running on the same server.

The project tree now contains a new process node named after the uploaded file. Through the process monitor you are notified as soon as the data transfer has finished. Once the data loading is completed, the process node will turn into a gene set node.

Figure 3.3 Gene data set in the project tree



Node statistics (number of genes, TRANSFAC matrices, promoters and molecules, respectively) as well as the creation time are shown near the name. Display of this additional information can be adjusted via the "Preferences" dialog item (see Section 1.5.9). Please note that the number of genes, matrices, promoters and molecules are not identical. In this example, for 126 genes there are:

20 matrices - Able to bind transcription factors from the list of 126 proteins

244 promoters - Note: A gene might have more than one promoter

69 molecules - These 69 out of the 126 molecules in the set are known to play a role in signal transduction pathways.

3.1.3 New gene set dialog

When you click on "Gene set" in the "Create new data" section of the ExPlain "File" menu, a "New gene set" dialog window will open. In the text field of this dialog you can type or paste identifiers to load them as a gene set. Note that this form is for identifiers only, so anything unrecognizable will be discarded. If you want to add other data (like expression values), please create a tab-separated file and load the data as a file. You can also specify the database and species of your accession numbers, set the destination folder, and type a name for the gene set.

3.1.4 Import data

ExPlain allows the import of predefined system data sets as well as sets from other sources through the "Import data" dialog of the "File" menu. In the upper part of the dialog you can specify the folder into which the files will be imported. Below this option you can see the lists of available data.

Note that the Import dialog allows multiple selections from both lists (using Shift and Ctrl keys), so you can import several data sets simultaneously.

ExPlain provides a number of installed gene sets, interval data, and profiles. Note: Human, Mouse, and Rat promoters and gene sets are not available in ExPlain-Plant, while Arabidopsis, Rice, and Soybean promoters are available exclusively in the ExPlain-Plant system.

For example, to upload the "HUVEC GSE2639 example" select it in the ExPlain predefined data (left) list of "Import data" dialog window, and press the OK button. This will launch a process to load the genes. It may take some time to load the data.

ExPlain predefined profiles (you can read more on profiles in Chapter 8, «Profiles») also appear in the same list and can be imported. These profiles are usually imported by default on the first run of ExPlain and later can be removed or modified. If you need a clean version of one of the system profiles, just import it again.

Figure 3.4 New gene set dialog

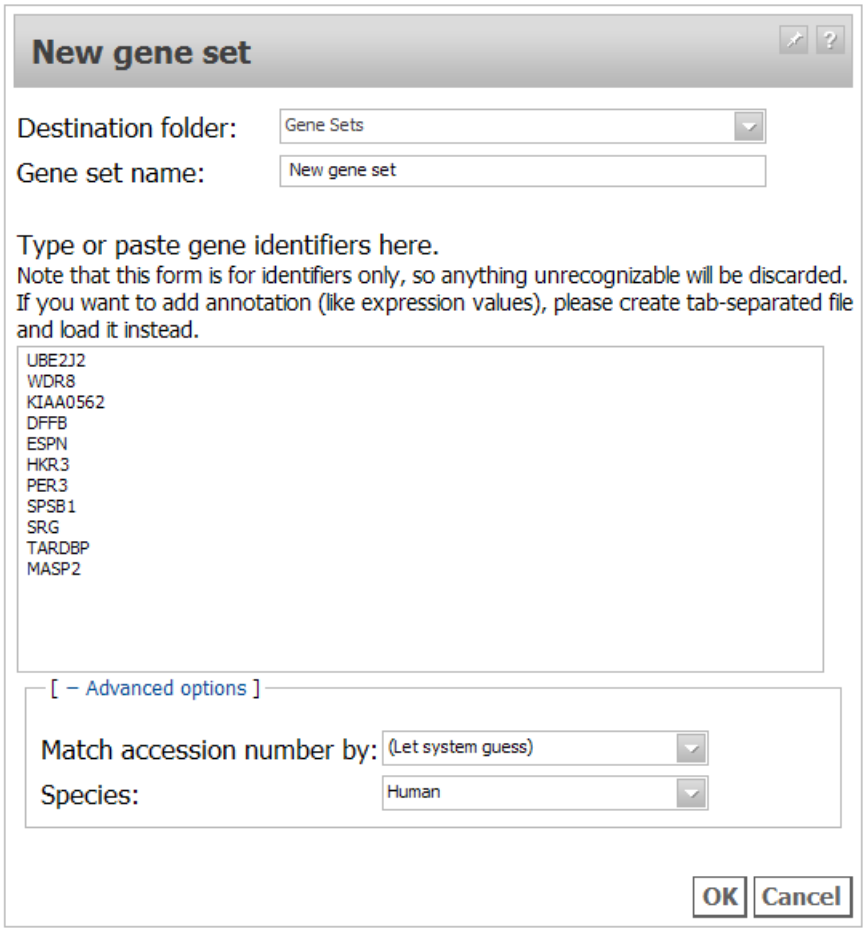
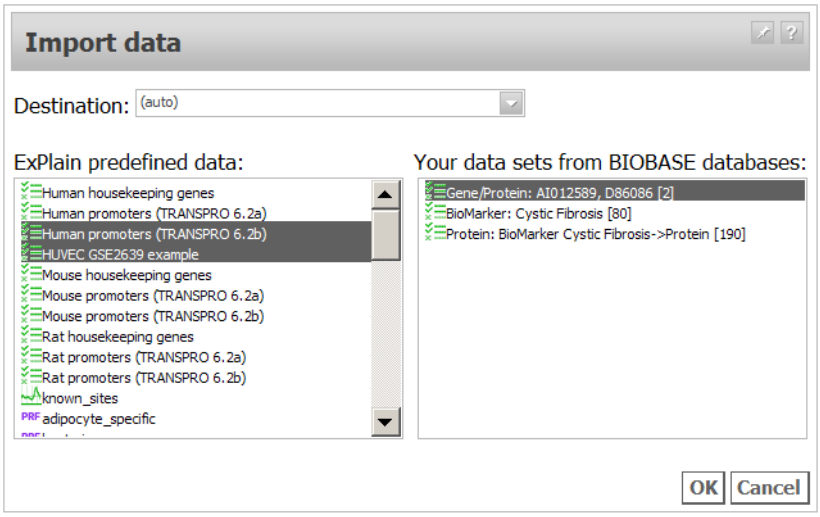


Figure 3.5 Import data dialog with the saved BKL search results



The data set "Known sites" is the set of genome intervals that are experimentally proved to correspond to real sites from TRANSFAC database.

The right side list in the "Import data" dialog window contains various data sets from BIOBASE GmbH databases. For example, if ExPlain is installed together with the BKL database, you can select certain saved search results from BKL to be imported as an ExPlain gene set. In the right menu you can see the names of the saved results that can be imported (Figure 3.5). The number of rows in each search

result is shown in the square brackets next to its name.

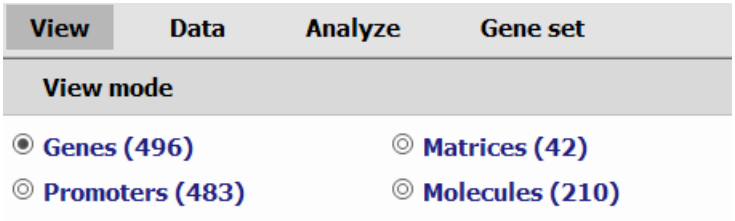
User-defined profiles from TRANSFAC installation are also displayed in the right side list. For a description of how to create profiles in TRANSFAC Professional, e.g. on the basis of a database search result, please consult the documentation for TRANSFAC and MATCH.

3.2 Representation of the gene set data

The data table is shown in the workspace by clicking its node link in the project tree. If more than 25% of the loaded gene identifiers were not matched by ExPlain, a warning message will appear at the top of the gene set page, stating how many of your identifiers could not be matched by the system. In the gene set table next to the checkbox column, ExPlain has added several columns which were not present in the original data table. These are system annotation columns which are taken from the ExPlain database and provide additional information to your original set. Some of them are hidden by default, but can be made visible via the "show columns" control or "Manage columns in data set" dialog (see Section 1.5.5).

Entity matching information is given in the "View" menu, showing number of genes, matrices, promoters and molecules.

Figure 3.6 Entity information in "View" menu



NOTE

Please note that the number of genes, matrices, promoters, and molecules are not identical. In the example (Figure 3.6), for 496 genes there are: 42 matrices that are able to bind transcription factors from the list of 496 proteins, 483 promoters - since one gene might have more than one promoter, and 210 molecules (out of 456) that are known to play a role in the signal transduction pathways.

To change the view of your gene set and see all molecules connected with your data click on the corresponding link. The number of rows you see is exactly the number of corresponding entities for the current view. If your data set has 100 genes, you will see exactly 100 rows in the table even if you had a different annotation for the same gene in source data file. In this case corresponding annotation fields are merged into effective values. Effective values are those passed into analyses and used for sorting in the table. These effective values can be manipulated. For example if you have two expression values for the same gene, you can specify whether the effective value is the minimum, maximum or average of the source expression values. In Figure 3.7 you can see an example of such case. A gene with the symbol Abcc9 has two corresponding accessions in the initial file, thus two different fold-change values. The effective value used for sorting is by default the average, but this can be changed in column preferences. The content of the text, dblink and categorical columns is also modified to fit the current number of rows. All column settings for the current set can be controlled via the "Rename/customize column" dialog from the "Data" menu (Section 1.5.6).

Press the [+] link near the gene set name to see additional information on the file origin. ExPlain allows annotation of projects through project notes as well as storage of technical conditions of the microarray experiment extracted from the respective Affymetrix™ DTT file. For more information on the data table functionality, columns management, subset renaming etc. see Section 1.5.

The "Assign levels to columns" option of the "Data" menu launches the Factor and Level assignment dialog described in Section 11.1.1. You can group dataset columns by levels within a category referred to as "factor". After the assignment is done, it is displayed as additional lines above the data table (two

Figure 3.7 Gene set representation

[+] **Gene set:** HUVEC GSE2639 example

Species name: **Human**

Filter (filter bar): none (total 7985 rows)

Rows per page: 50

Export: Plain text | XLS | RTF

1 2 3 4 5 6 7

Mark: Page (50) | All (7985) | None | Invert

	Gene symbol	BKL description	Fold change
(*)	A2M	Alpha-2-macroglobulin; binds to collectin; plays a role in cell proliferation and protein homotetramerization; upregulated in Alzheimer disease; sickle cell anemia; rheumatoid arthritis; multiple sclerosis; and prostatic neoplasms	1.02
(*)	A4GNT	Alpha-1;4-N-Acetylglucosaminyltransferase; a glycosyltransferase that forms alpha 1;4-linked GlcNAc residues; especially in O-glycans; and is involved in synthesis of class III mucins	1.05
(*)	AADAC	Arylacetamide deacetylase; may play a role in protein amino acid deacetylation and lipid metabolic process	1.01
(*)	AAK1	AP2 associated kinase 1; a protein serine-threonine kinase that acts in protein amino acid phosphorylation and protein import; regulates receptor-mediated endocytosis	1.01
(*)	AANAT	Arylalkylamine N-acetyltransferase; acts in melatonin biosynthesis; involved in response to xenobiotic stimulus; increased mRNA expression correlates with basal cell carcinoma; single nucleotide polymorphism is associated with idiopathic scoliosis	1.03
(*)	AARS	Alanyl-tRNA synthetase; a tRNA binding protein that is involved in humoral immune response and tRNA processing; acts as an autoantigen in dermatomyositis	1.01

Figure 3.8 Gene set annotation window

[−] **Gene set:** HUVEC GSE2639 example

Origin
Gene set was loaded from predefined set "HUVEC GSE2639 example". Of 11840 accessions 10035 were recognized and matched. The rest 1805 accessions were ignored.

Data build
Mammal.2007-11-21

User comments

Attach descriptions from Affymetrix® DTT XML file or ARR file: Browse...

Update

factors ”Time factor” and ”Temperature” are added to a dataset on the figure below). To remove or change the assignment, launch the dialog again, remove unnecessary factors and/or add new ones.

Figure 3.9 Gene set with assigned factors

Time factor »		1 h	2 h	1 h		
Temperature »		low	medium	high		
	Gene symbol	H02.CEL	H11.CEL	H24.CEL	H03.CEL	H12.CEL
(*)	ACO1	8.29	3.85	4.79	7.42	2.9
(*)	ACTB	44.4	39.3	39.4	42.6	47.5
(*)	ACTG1	22	21.8	19.5	16.7	17.5

3.3 Recombining gene sets

Syntax of naming recombined nodes. By default, node names in the project tree convey information about the way they were obtained. For all sets created within a corresponding project, by default the name is prefixed with a number string, where number increments chronologically starting with 1, and carries the number of entities in parentheses and a timestamp. Names of recombined sets are constructed from the names of the nodes involved and the applied operation. Here Sx&Sy signifies a set created by union, Sx|Sy is a set that was created by intersection and Sx\Sy is a set that was created by subtracting Sy from Sx.

3.3.1 Filter gene set by condition

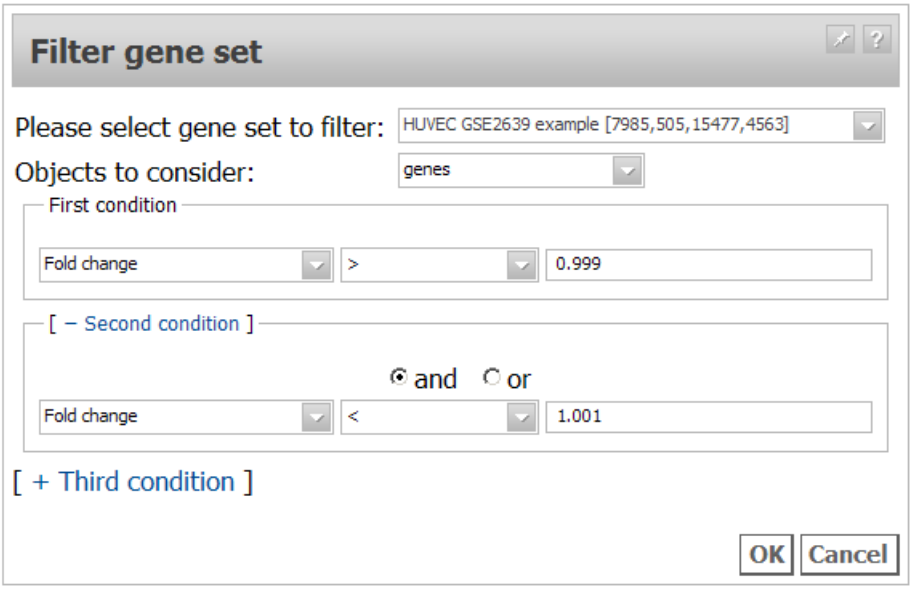
The ”Filter gene set by condition” dialog of the ”Data” menu can be used to extract data entities based on a range of expression, numerical values or text values as well as distinct categories. This allows you to easily design experiments that compare different combinations of expression levels or changes.

NOTE

This tool is not limited to a single column. You can choose up to three different columns with specific conditions on each of them.

Figure 3.10 shows the ”Filter gene set by condition” dialog with configured parameters. First you should select the gene set that will be filtered. The field ”Objects to consider” allows you to select the entities which will be used to create the filtered set: genes, promoters, molecules or matrices. Note, for example, if you select the ”matrices” view, then genes without a corresponding position weight matrix will not appear in the result set even if they satisfy the conditions. Think of it as switching the gene set table to the corresponding view and then manually selecting some lines by checking checkboxes in the table. If you are not sure about this option, use ”genes”. The filtering parameters can be set up using condition fields. The second and the third fields are collapsed by default. To expand other conditions, click the corresponding links. Several conditions can be connected by ”and” or ”or” rules. You can use up to three columns to specify the entities you wish to extract.

Figure 3.10 Subset creation with expression value constraints



Each condition consists of three fields. The column list contains names of columns of the selected set, regardless of their type and visibility. Note that columns with system annotation are marked in the list with the picture and other columns have the picture on the left. Subsequent lists are used to set a requirement for a marked column. For a text or db-link column you can require entities to contain, start or end with a particular string. For a numerical expression column, you can require corresponding entries to be equal, lower, higher, or not equal to the specified value. Categorical values can be required to be equal or not equal to the specified category.

As an example, we want to extract non-changed genes from the ”HUVEC GSE2639 example” predefined dataset. We select the expression column named ”Fold change” in the first and in the second condition fields. We seek all entities that have fold change greater than 0.999 AND lower than 1.001 (Figure 3.10). Subset creation is invoked with the OK button.

After the process is done we have a new subset with nonchanged genes placed under the original

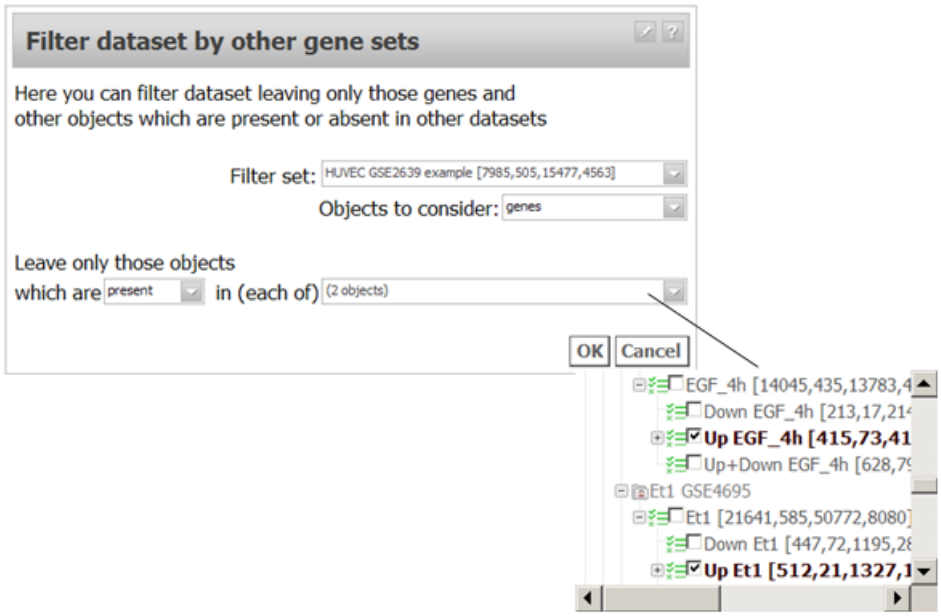
set in the project tree as it is shown at Figure 3.11. The new set contains 777 genes.

Figure 3.11 Subset created with expression value constraints

3.3.2 Filter gene set by other gene sets

In ExPlain you can filter a geneset leaving only those genes and other objects which are present or absent in some other datasets. Choose the "Filter by gene sets" option in the "Data" menu. The active tree node will be selected by default in the tree-control lists. Select the set to be filtered and the other sets by choosing "absent" or "present" to perform subtraction or intersection respectively, specify objects to consider, and press the OK button.

Figure 3.12 Filter geneset by another geneset dialog



3.3.3 Join two gene sets

Similarly you can create a new set with a unique list of elements from several source sets. Open the "Join gene sets" dialog window from the "Data" menu. A dialog window with the subsets selected for joining is shown in the Figure 3.13. If gene sets were prepared from the same original data set, you will have the same number of columns in the resulting set. When joining sets of different origin check "Merge native columns: When name and type equals" to join columns with the same name and type, or leave it unchecked to have all columns distinct.

After pressing the OK button, ExPlain creates the recombined set, as a child node of the first gene set.

3.3.4 Extracting unique genes

You can select several genesets to create subsets of genes, which are unique within this group (i.e. present in only one set, but absent in any other selected gene set). Open the "Extract unique genes" dialog window from the "Data" menu, specify objects to consider, selecting at least two items, then press the OK button (Figure 3.14).

Figure 3.13 Join two gene sets dialog

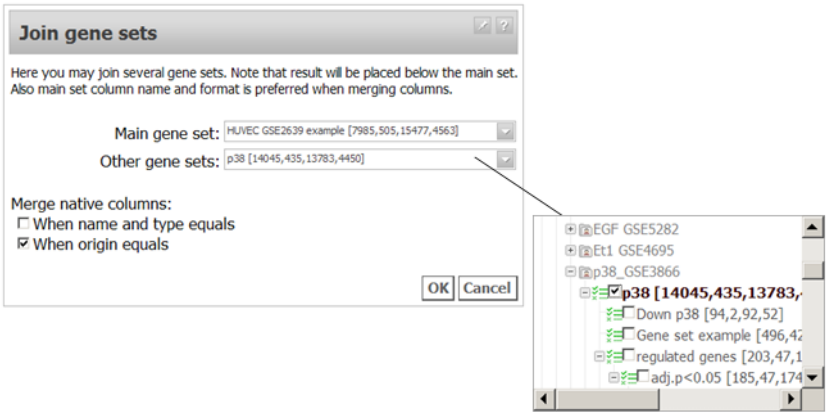
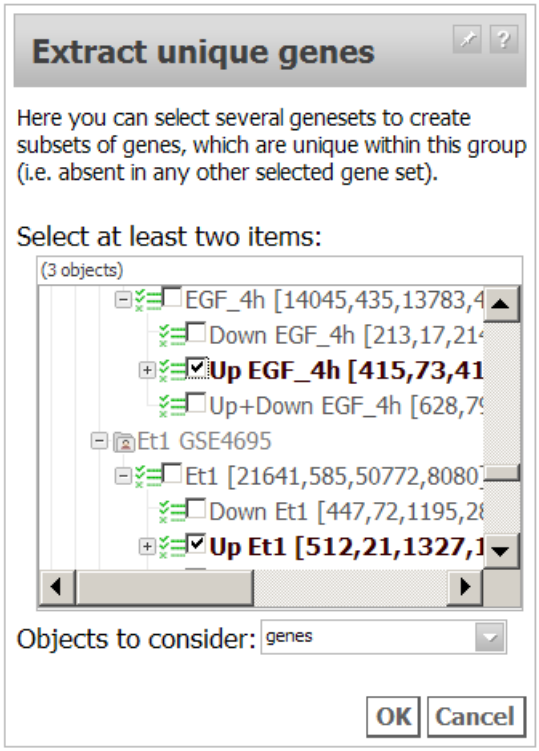


Figure 3.14 Extract unique genes dialog

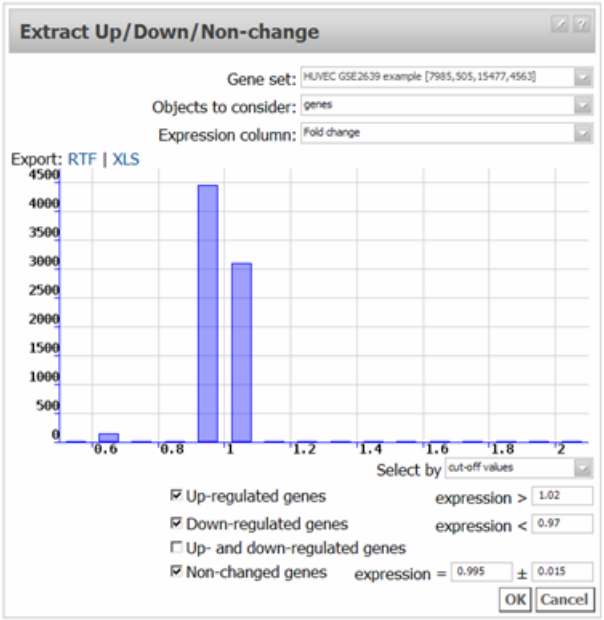


3.3.5 Extract Up/Down/Non-change

You can easily create from your gene set up to four subsets representing up-regulated genes, down-regulated genes, genes with non-changed expression and genes with changed expression (both up and down in the same set). You may perform the same action using 'Filter gene set by condition' dialog several times for each resulting subset, but 'Extract Up/Down/Non-change' does this in one single step. This dialog also displays the histogram of expression values distribution, so you can use it just to view this distribution without creating any sub-sets.

First select the gene set you want to create sub-sets from and the objects that will be used as described above in Section 3.3.1. Next specify the column which will be considered as the expression column. After that you'll see the histogram which reflects the distribution of expression values. The X axis represents expression value, while the Y axis represents the number of objects in the selected view (number of genes if you selected 'genes' before), having an expression value lying in the specific interval. The number of bars as well as bar width is selected automatically. Also, up to 1% of minimal and 1% of maximal points may not be displayed if they are too far from others. For example, if you have 99 genes

Figure 3.15 Extract Up/Down/Non-change dialog



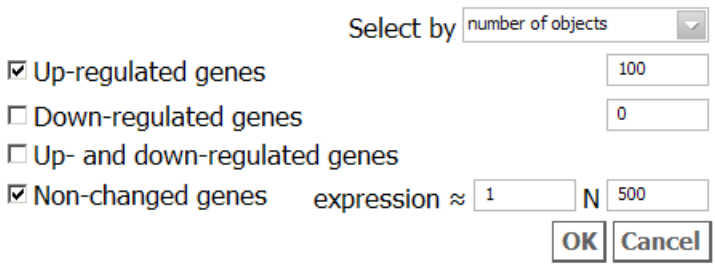
with expression between 0 and 10 and one gene with expression value 20, then the graph will be scaled to display expression values between 0 and 10 only.

Below the graph you can specify which sub-sets you want to create by using the checkboxes. At the right side you can define how the genes (or other objects) will be distributed among sub-sets. There are two ways to define this: either specifying cut-off values or inserting the number of objects that will be selected. In the 'cut-off values' mode you can specify the minimal expression value for up-regulated sub-set, the maximal expression value for the down-regulated sub-set and mean value and range of expression values for the non-change sub-set.

In the 'number of objects' mode (see picture below) you can specify the number of genes (or other objects) which will be included in the up (NU), down (ND) and non-change (NNC) sub-sets. For the non-change sub-set you should also specify the mean expression value (ENC). Thus, the up-regulated sub-set will contain exactly NU genes (or other objects) with the highest expression, the down-regulated sub-set will contain exactly ND objects with the lowest expression and the non-change sub-set will contain exactly NNC objects with the expression values closest to ENC.

Note that if you have several genes with exactly the same expression value, the result may be unpredictable. For example, if you have gene A with expression value 3, gene B with expression value 2 and gene C with expression value 2 and set NU = 2, then gene A will appear in the up-regulated sub-set for sure, and one of genes B and C will appear also, but you cannot predict which one. Thus this mode is less predictable than the 'cut-off values' mode, but still, if you specify the same parameters on the same gene set twice, it's guaranteed that the result will be the same even in 'number of objects' mode.

Figure 3.16 Dialog parameters in 'number of objects' mode

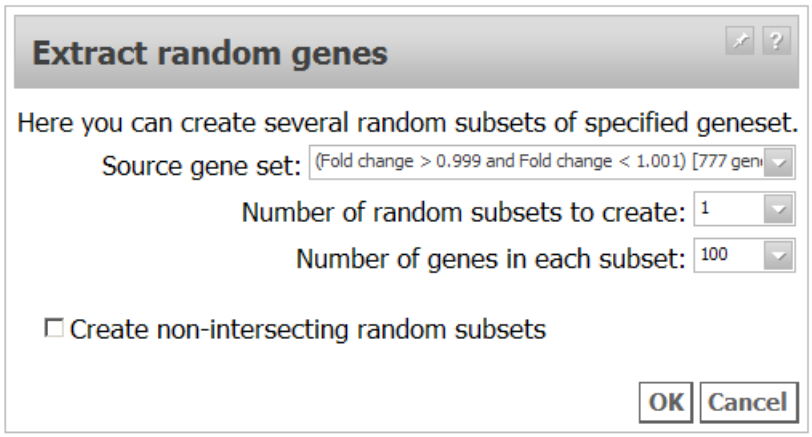


In both modes, 'Up- and down-regulated genes' is just the combination of Up- and Down- sub sets.

3.3.6 Extracting random subsets

You can you can create several random subsets using any geneset in the tree. Open the "Random subsets" dialog window from the "Data" menu, and specify the source gene set, number of random subsets to create, number of genes in each subset, and then press the OK button (Figure 3.17). If the "non-intersecting" option is checked, the resulting subsets will contain different sets of genes.

Figure 3.17 Extract random subsets dialog


Extract random genes
Here you can create several random subsets of specified geneset.
Source gene set:
(Fold change > 0.999 and Fold change < 1.001) [777 gen
Number of random subsets to create:
1
Number of genes in each subset:
100
Create non-intersecting random subsets
OK
Cancel

3.4 Adding columns to the gene set

3.4.1 Calculation from an existing numerical column

For analyses that take expression values into account, it can be useful to transform the values beforehand. Logarithms, for instance, can be used to decrease differences between fold changes. Multiplication by -1 or absolute values can alter the way genes with negative fold changes are taken into account. Some statistical methods such as normalization and mean shifting can also be necessary for data processing. In a case where multiple numerical columns exist, the average values, sum or multiplication of values in different columns, is sometimes meaningful.

ExPlain provides a range of possibilities for such transformations of numerical and expression column values, summarized below.

Arithmetic operations + (add), - (subtract), * (multiply), / (divide), ^ (power).

Advanced operations abs(x) (absolute value), sqrt(x) (square root), exp(x) (exponentiation), log(x) (natural logarithm), log2(x) (binary logarithm), log10(x) (logarithm to the base 10)

Trigonometrical functions sin(x) (Sine), cos(x) (Cosine)

Aggregate functions sum(x) (summ of all values in the column), avg(x) (average value in the column), max(x) (maximal value in the column), min(x) (minimal value in the column)

Parenthesis "(" and ")" parenthesis are used to group parts of the formula

Columns suitable for calculations are listed in the "Available columns" control of the "Calculate column" dialog that can be called via the "Compute from existing columns" item of the "Data" menu. To use a column in a formula you can print a column denotation (the "#" sign and the number), or you can click on the column name in the list. The corresponding column symbol will then appear in the formula line. The formula should consist of column names, numbers, and functions (listed above), with the

symbol ”x” replaced with a column denotation number. Numbers can be in normal format (123.456) of in scientific format (1.23456e+02).

In the example shown in Figure 3.18 we selected the Fold change column (denoted as #1) and apply a logarithmic transformation to absolute values of the column. As a result we obtain a dataset with the original column (Fold change) and a new one (log2(|Fold change|) with transformed fold changes. Figure 3.19 shows a part of the data table with the newly added column.

Figure 3.18 Calculation from existing numerical column dialog

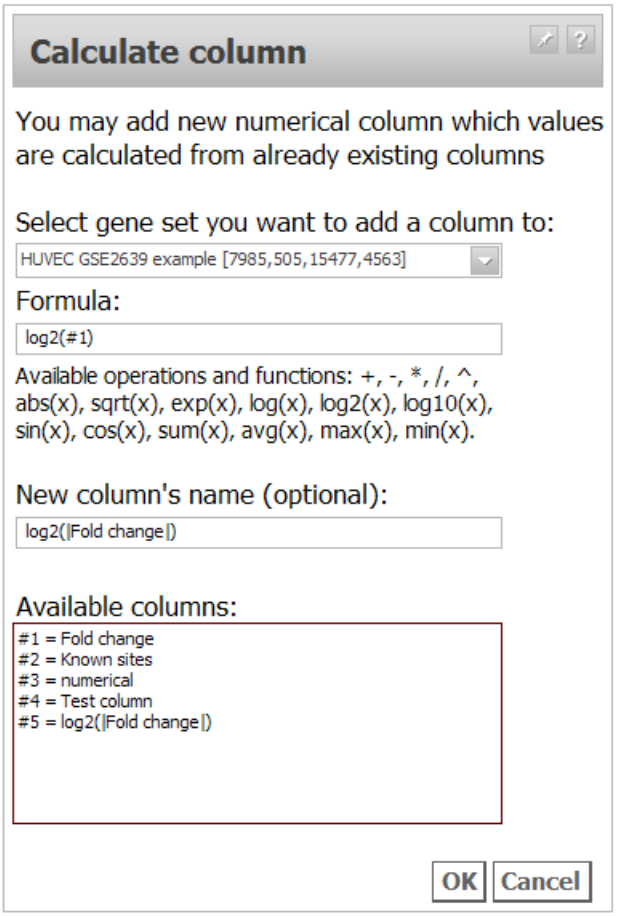

Calculate column
You may add new numerical column which values are calculated from already existing columns
Select gene set you want to add a column to:
HUVEC GSE2639 example [7985,505,15477,4563]
Formula:
log2(#1)
Available operations and functions: +, -, *, /, ^, abs(x), sqrt(x), exp(x), log(x), log2(x), log10(x), sin(x), cos(x), sum(x), avg(x), max(x), min(x).
New column's name (optional):
log2(|Fold change|)
Available columns:
#1 = Fold change
#2 = Known sites
#3 = numerical
#4 = Test column
#5 = log2(|Fold change|)
OK
Cancel

Figure 3.19 Data table with logarithmically transformed of a fold change column

Gene symbol	BKL description	Fold change	log2(|Fold change|)
A2M	Alpha-2-macroglobulin; binds to collectin; plays a role in cell proliferation and protein homotetramerization; upregulated in Alzheimer disease; sickle cell anemia; rheumatoid arthritis; multiple sclerosis; and prostatic neoplasms	1.02	0.0285692
A4GNT	Alpha-1;4-N-Acetylglucosaminyltransferase; a glycosyltransferase that forms alpha 1;4-linked GlcNAc residues; especially in O-glycans; and is involved in synthesis of class III mucins	1.05	0.0703893
AADAC	Arylacetamide deacetylase; may play a role in protein amino acid deacetylation and lipid metabolic process	1.01	0.0143553
AAK1	AP2 associated kinase 1; a protein serine-threonine kinase that acts in protein amino acid phosphorylation and protein import; regulates receptor-mediated endocytosis	1.01	0.0143553
AANAT	Arylalkylamine N-acetyltransferase; acts in melatonin biosynthesis; involved in response to xenobiotic stimulus; increased mRNA expression correlates with basal cell carcinoma; single nucleotide polymorphism is associated with idiopathic scoliosis	1.03	0.0426443
AARS	Alanyl-tRNA synthetase; a tRNA binding protein that is involved in humoral immune response and tRNA processing; acts as an autoantigen in dermatomyositis	1.01	0.0143553

3.4.2 Linking a column from another data set

In order to compare different results you can add to your data set columns from other data sets. In the example shown on Figure 3.20 we will add to the set of HUVEC genes the ”TRANSPro accession” and ”Gene symbol” columns from the Human Promoters set. It’s also possible to link columns from different analyses’ results. Choose the ”Link from another data set” option from the ”Add columns” section of

the "Data" menu. When you select a dataset with the columns that you want to add, the form will be disabled while system refreshes the list of available columns. Then select column(s) and press the OK button. If you would like to link columns to more than one gene set, select "Multi-select mode" in the "Gene set you want to add a column to" menu. Linked columns will be marked by the icon. Click on it to go to the original dataset. You can change the format of the linked column, rename it, and sort or filter its data.

Figure 3.20 Link column from another data set dialog

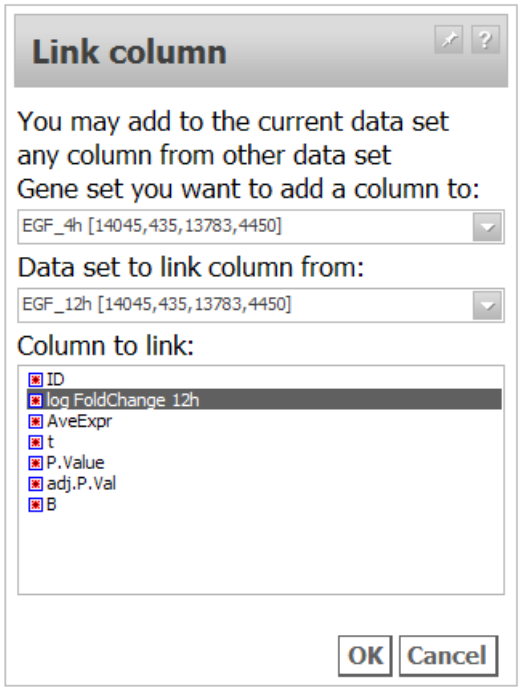


Figure 3.21 Gene set with linked column

	Gene symbol	BKL description	log FoldChange 4h ▼	log FoldChange 12h
(*)	Serpinb2	Plasminogen activator inhibitor 2; a putative endopeptidase inhibitor that plays a role in wound healing; human SERPINB2 is associated with several types of neoplasms	9.00057	6.19307
(*)	Hmga1	High mobility group AT-hook 1; a transcriptional regulator that functions in cell proliferation and DNA repair; human HMGA1 is associated with leiomyoma; lung; breast; and other various cancers; mouse Hmga1 is associated with cardiac hypertrophy	8.69926	6.45499
(*)	Fosl1	Fos like antigen 1; a transcription activator; acts in neuron differentiation; gene expression is upregulated in ischemia and asbestos induced mesothelioma; human FOSL1 is associated with several cancers; mouse Fosl1 is associated with osteopenia	8.53336	5.9523
(*)	Areg	Amphiregulin; an EGFR activating ligand that acts in antiapoptosis; chemotaxis; cell proliferation; prostaglandin biosynthesis; and wound healing; human AREG is associated with asthma; multiple myeloma; HIV infection; psoriasis; and several neoplasms	8.42203	6.41408
(*)	Aldh1a3	Aldehyde dehydrogenase 1 family member A3; plays a role in vitamin A metabolism; brain development; and camera-type eye morphogenesis; may act in generation of neurons and alcohol metabolism; human ALDH1A3 is associated with psoriasis	7.26606	0.841571
(*)	Dusp5	Dual specificity phosphatase 5; a phosphoprotein phosphatase that acts in cytokine-mediated and glutamate signaling pathway; inhibits MAPK activity; involved in response to heat and stress; human DUSP5 is associated with thyroid neoplasm	7.11151	5.70001

3.4.3 Adding a column with system annotations

ExPlain provides annotation columns containing some additional information such as identifiers of external and internal databases, information about molecule sub- or superfamilies, data from the BioKnowledge Library (BKL) database about genes, molecules and other biological objects. This information can be attached to your gene sets. To do so choose the "System annotation" option from the "Add columns" section of the "Data" menu, then select an appropriate gene set, choose one or several columns from the list, and press the OK button. If you would like to add columns with system annotation to more than one gene set, select "Multi-select mode" in the "Add column(s) to gene set" menu (Figure 3.22).

Figure 3.22 Add columns with system annotation dialog

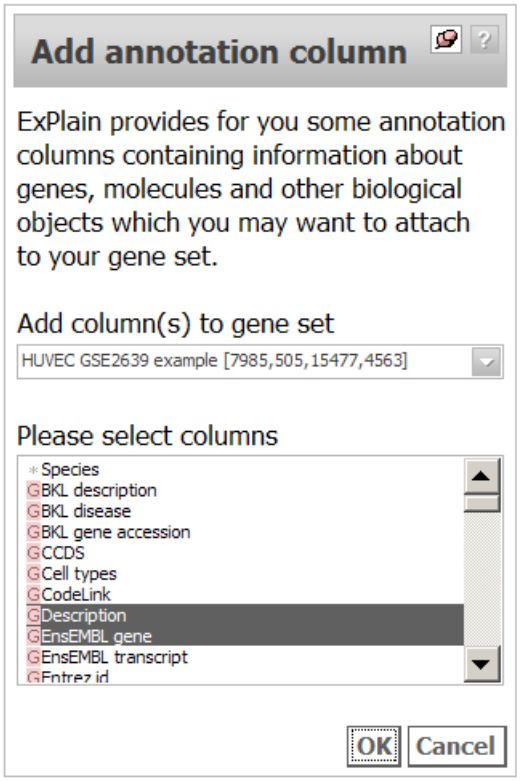


Figure 3.23 Gene set with two annotation columns

Gene symbol	Description	EnsEMBL gene	Fold change
A2M	alpha-2-macroglobulin	ENSG00000175899	1.02
A4GNT	alpha-1,4-N-acetylglucosaminyltransferase	ENSG00000118017	1.05
AADAC	arylacetamide deacetylase (esterase)	ENSG00000114771	1.01
AAK1	AP2 associated kinase 1	ENSG00000115977	1.01
AANAT	arylalkylamine N-acetyltransferase	ENSG00000129673	1.03
AARS	alanyl-tRNA synthetase	ENSG00000090861	1.01
AARSD1	alanyl-tRNA synthetase domain containing 1	ENSG00000108825	0.96
AATK	apoptosis-associated tyrosine kinase	ENSG00000181409	0.97
ABCA1	ATP-binding cassette, sub-family A (ABC1), member 1	ENSG00000165029	1.04
ABCA11P	ATP-binding cassette, sub-family A (ABC1), member 11 (pseudogene)	ENSG00000186777	1.04

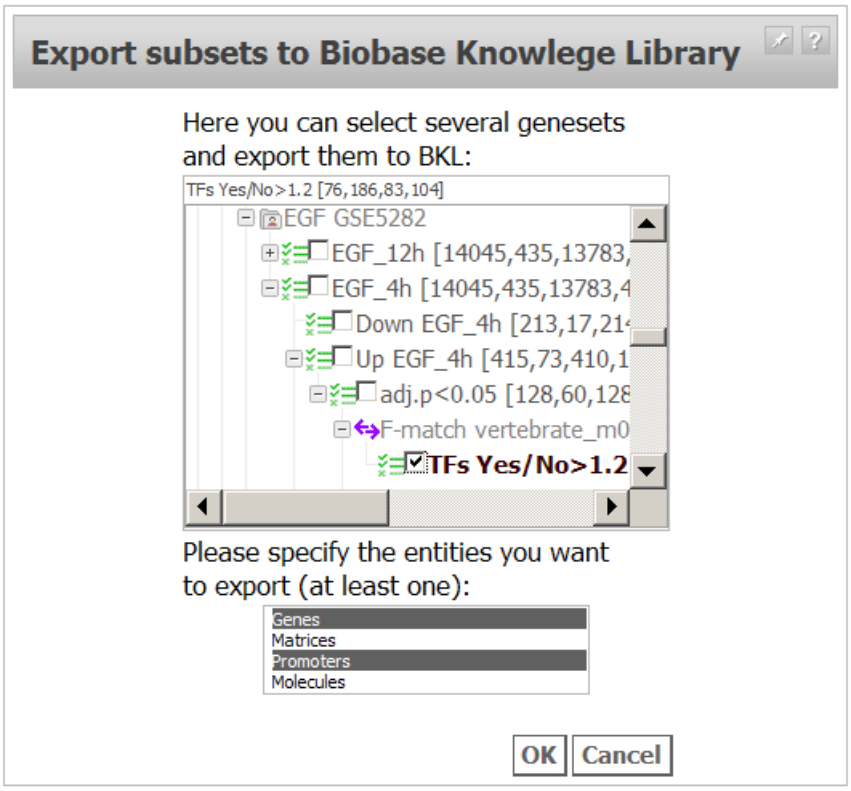
3.5 Exporting gene sets to BKL

You can export gene sets from ExPlain to the BioKnowledge Library (BKL). To do so choose the ”Export gene sets to BKL” option from the ”File” menu, then select one or more gene set to be exported, specify at least one entity from the list, and press the OK button (Figure 3.24). All exported sets will appear in your Search Results in the User data section of the BKL page.

3.5.1 Exporting selected genes as BKL search result

You can export selected genes from ExPlain to BioKnowledge Library (BKL) as BKL search results. To do so click on the appropriate gene set in the three, mark genes you want to export, and then choose the ”BKL search result” option from the ”Gene set” menu. All exported genes will appear in your Search Results in the User data section of the BKL page.

Figure 3.24 Exporting gene sets to BKL dialog



3.6 Statistics calculator

You can use the integrated statistics calculator to run classical statistical methods on your data. To do so, choose the ”Statistics calculator” option from the ”Analyze” menu. Then in the dialog window, specify the gene set to analyze, objects to consider, and two numerical columns to be compared. The figure below shows an example of such an analysis.

Figure 3.25 Statistics calculator dialog window

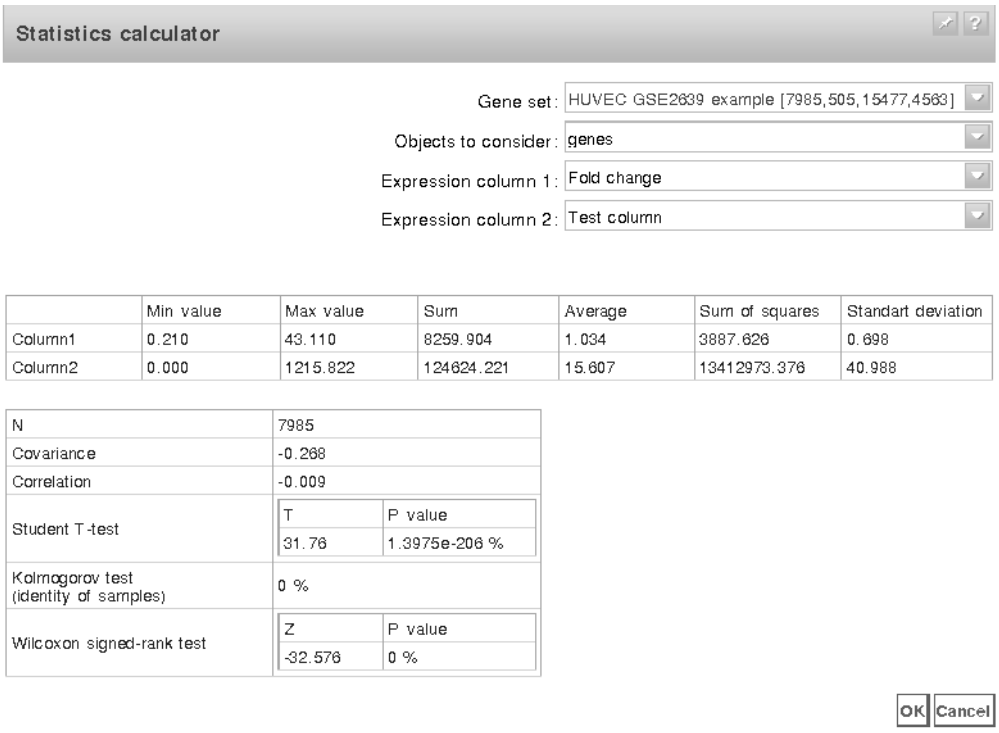


Table 3.2 Predefined data sets

Set name	Description
Human housekeeping genes	This set contains about 550 human housekeeping genes derived from a study described in [8].
Human promoters TRANSPro 6.2	This set contains over 40,000 human promoters extracted from TRANSPro.
HUVEC GSE2639 example	This set contains over 7,500 human genes with an associated fold change value. Human umbilical vein wall's cells were treated with TNF-alpha (tumor necrosis factor alpha) and a microarray experiment was done. the fold change is the ratio of signals of treated cell genes and control cell genes.
Mouse housekeeping genes	This set contains over 400 mouse housekeeping genes.
Mouse promoters TRANSPro 6.2	This set contains about 61,000 mouse promoters from TRANSPro extracted with the same conditions as the human promoter set.
Rat housekeeping genes	This set contains about 400 rat housekeeping genes.
Rat promoters TRANSPro 6.2	This set contains over 21,000 rat promoters from TRANSPro extracted with the same conditions as the human promoter set.
ExPlain-Plant only	
Arabidopsis promoters TRANSPro 6.2	Available in ExPlain-Plant, this set contains over 27,000 Arabidopsis promoters from TRANSPro.
Rice promoters TRANSPro 6.2	Available in ExPlain-Plant, this set contains over 29,000 rice promoters from TRANSPro extracted with the same conditions as the Arabidopsis promoter set.
Soybean promoters TRANSPro 6.2	Available in ExPlain-Plant, this set contains over 68,000 soybean promoters from TRANSPro.

Chapter 4

The Functional classification

This chapter describes the Functional Analysis (FA) and Gene Set Enrichment Analysis (GSEA) algorithms.

The FA module allows you to identify statistically relevant calssification terms, including Gene Ontology [22], disease associations and specific organ/tissue expression annotations.

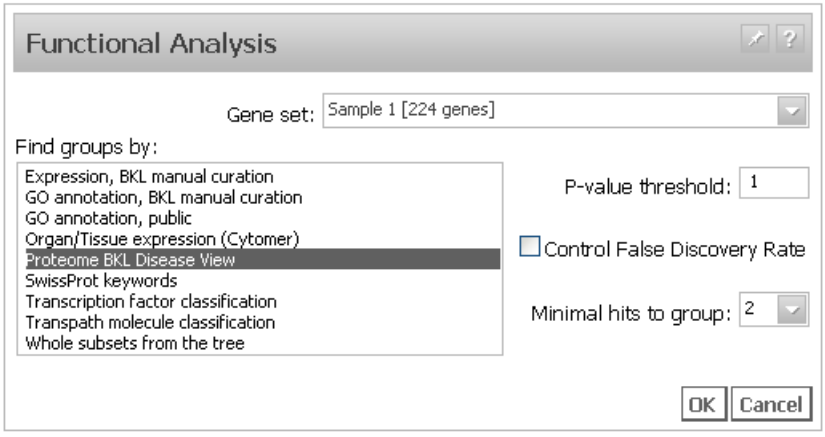
The GSEA module enables you to explore statistical enrichment of functional annotations in your gene sets taking into account the expression value, fold change or other numerical column associated to them.

These analyses allow you to extract subsets of functionally related genes, and to investigate the biological properties of any gene list. Furthermore, you can use these algorithms to compare a given set to other data sets on your project tree. If you have a list of genes compiled for your specific research interest, you can load it into ExPlain and extract genes of this list from any other gene set.

4.1 The Functional classification analysis

4.1.1 How to run classificaiton analysis

Figure 4.1 Dialog window of the Functional classification



- Click the "Functional classification..." menu link in the "Analysis" menu or press to launch the dialog window.

- Select analyzed set.

- Select functional category or categories.

- Set thresholds for P-value and minimal number of hits.

- Press "OK" button.

- To extract a list of genes assigned to functional groups select the groups and press icon.

Species in the Functional categories (human, mouse, rat) are defined automatically.

4.1.2 Functional Analysis categories

Expression, BKL manual curation The Annotated groups here are different organs, tissues, tumors or cell types of the human organism. Mouse and rat genes are abstracted to ortholog level to be mapped.

GO annotation, BKL manual curation GO groups are terms of the respective Gene Ontology hierarchy, manually curated in the BKL

GO annotation, public Gene Ontology hierarchy from files submitted by GO Consortium members to Gene Ontology <http://www.geneontology.org>

Organ/Tissue expression (Cytomer) Genes associated to different organs or tissues of the human organism or, in the case of mouse and rat genes, matched to the closest ortholog.

Proteome BKL Disease View Genes are associated to diseases annotated in the BKL database. Genes are classified as connected to diseases in correlative, preventative or causal relationships. For mouse and rat they are matched to corresponding orthologs.

SwissProt keywords Genes are matched to terms from the UniProt Knowledgebase that include widely accepted biological ontologies, classifications, and cross-references.

Transcription Factor Classification Genes are mapped to a classification of transcription factors (human, mouse or rat) annotated in the TRANSFAC Professional database.

TRANSPATH Molecule Classification Genes are mapped to a classification of molecules (human, mouse or rat) annotated in the TRANSPATH Professional database.

GRO - Plant Growth Stages (rice only) Genes are matched to terms of the Plant Growth Stages hierarchy.

PO - Plant Structure and Growth Stages (rice and Arabidopsis) Genes are matched to terms of the Plant Structure and Growth Stages hierarchy.

Trait Ontology (rice only) Genes are matched to terms of the Trait Ontology hierarchy

EO - Environment Ontology (rice only) Genes are matched to terms of the Environment Ontology

Whole subsets from the tree Genes are matched to other gene sets in your project tree.

4.1.3 Functional analysis dialog window

The Functional analysis dialog window is shown in Figure 4.1. It contains a list to select the grouping category and editors to specify P-value and minimal hits thresholds. The "Control False Discovery Rate" checkbox adds additional verification in the p-value calculation algorithm, based on a Bonferroni correction.

It is possible to select several categories simultaneously, for which individual processes will be put in the queue and launched one after another. When the algorithm is launched from the active data node, this node is selected in the "Gene set" field automatically, you can change it or add more gene sets to the selected one (by switching the project tree in the dialog to the multi-selection mode).

4.2 Example analysis with BKL curated GO annotation

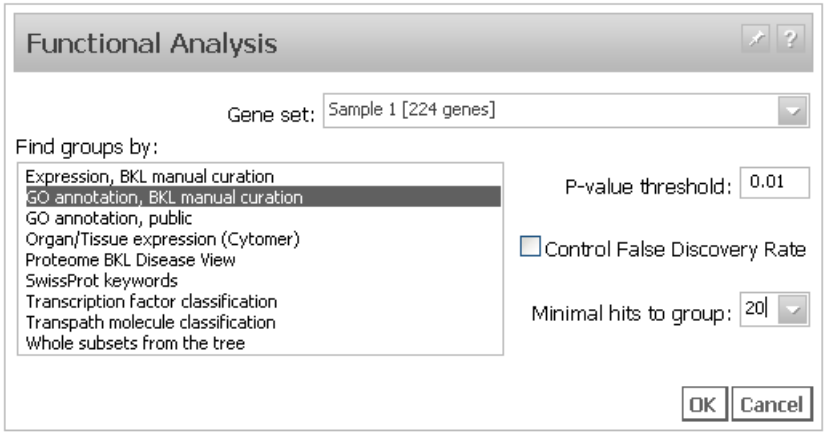
This example demonstrates how to use the Functional Analysis classification by analyzing a sample data set of human genes. We will use it to compile a subset of protein synthesis genes according to GO annotation curated by BKL.

The first step consists of selecting the desired data set from the project tree and navigating to the "Functional classification" button in the dropdown menu through the "Analyze" button in the main Menu. Alternatively, the "Classification ..." option can be chosen first, then the desired data set should be selected in the "Gene set" drop down menu in the FA dialog window. The desired category is selected from the category list. In our example we set the P-value threshold to 0.01 and the minimal number of hits to 20.

ASTUCE

The hit number threshold should be considered on a case-by-case basis. While larger hit numbers typically also correspond to more significant P-values, groups which are populated with fewer genes than those allowed by the threshold are missed. Thus, a high value is most useful for finding groups clearly enriched in the input list, while a vast collection of all groups of interest should be sought with a smaller threshold value.

Figure 4.2 Example of the Functional classification analysis

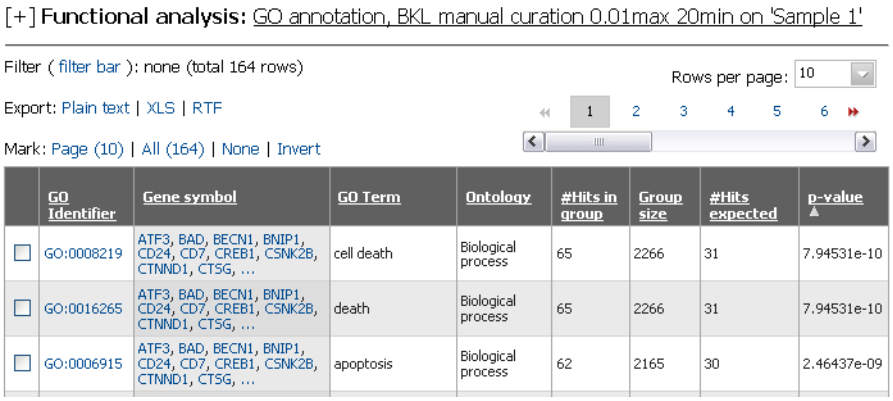


After clicking the OK button ExPlain adds a new node to the project tree under the active data set and displays the status of the process in the process monitor. Results can be inspected as soon as the analysis is done either by clicking on the link in the process monitor or in the project tree.

Matched terms of the selected Function category are given in each row of the output table. For GO ontologies, the table presents (from left to right) the GO identifier, Gene symbols, the GO term description, Ontology in general (Biological Process, Molecular Function or Cellular Component), the number of hits from the input set in the ontology group, the size of the group in the database, the

number of randomly expected hits, and the P-value of the observation. In our case, the cell death term is the most significant one (P=7.94531e-10).

Figure 4.3 Functional Analysis results with GO Biological Process groups



[+] **Functional analysis:** GO annotation, BKL manual curation 0.01max 20min on 'Sample 1'

Filter (filter bar): none (total 164 rows)

Rows per page: 10

Export: Plain text | XLS | RTF

1 2 3 4 5 6

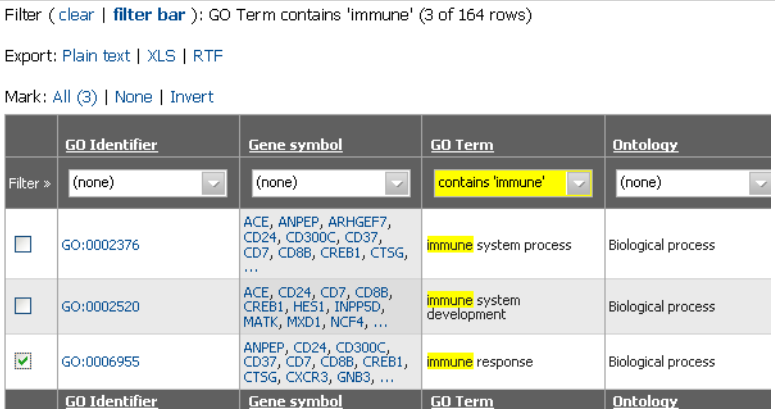
Mark: Page (10) | All (164) | None | Invert

	GO Identifier	Gene symbol	GO Term	Ontology	#Hits in group	Group size	#Hits expected	p-value
☐	GO:0008219	ATF3, BAD, BECN1, BNIP1, CD24, CD7, CREB1, CSNK2B, CTNND1, CTSG, ...	cell death	Biological process	65	2266	31	7.94531e-10
☐	GO:0016265	ATF3, BAD, BECN1, BNIP1, CD24, CD7, CREB1, CSNK2B, CTNND1, CTSG, ...	death	Biological process	65	2266	31	7.94531e-10
☐	GO:0006915	ATF3, BAD, BECN1, BNIP1, CD24, CD7, CREB1, CSNK2B, CTNND1, CTSG, ...	apoptosis	Biological process	62	2165	30	2.46437e-09

ASTUCE

Instead of scrolling the table manually, you can use the filtering option (see Section 1.5.3) to emphasize groups of interest. The figure below shows "immune"-highlited groups.

Figure 4.4 New subset with a defense, immune and inflammatory response genes compiled by Functional Analysis



Filter (clear | **filter bar**): GO Term contains 'immune' (3 of 164 rows)

Export: Plain text | XLS | RTF

Mark: All (3) | None | Invert

	GO Identifier	Gene symbol	GO Term	Ontology
Filter »	(none)	(none)	contains 'immune'	(none)
☐	GO:0002376	ACE, ANPEP, ARHGEF7, CD24, CD300C, CD37, CD7, CD8B, CREB1, CTSG, ...	immune system process	Biological process
☐	GO:0002520	ACE, CD24, CD7, CD8B, CREB1, HES1, INPP5D, MATK, MXD1, NCF4, ...	immune system development	Biological process
☑	GO:0006955	ANPEP, CD24, CD300C, CD37, CD7, CD8B, CREB1, CTSG, CXCR3, GNB3, ...	immune response	Biological process
	GO Identifier	Gene symbol	GO Term	Ontology

The available options of the "Functional classification" in the main menu are: "Gene set" , which compiles a new subset containing entries from all selected groups, and "Separate gene sets" , which creates an individual set for each group. New subset nodes will be added below the Function Analysis node in the project tree. The "GO Identifier" column with the ontology identifiers is hidden by default and can be shown if needed in the created subset.

4.3 Functional Analysis output tables

This section describes the output tables of the different Functional categories: Organ/Tissue expression, BKL Disease, Transcription Factor Classification, TRANSPATH Molecule Classification, Gene Ontology, and Whole subsets from the tree results.

Note, that only overrepresented groups are displayed in the result table for all types of classification.

Figure 4.5 New subset with an immune response and immune system development genes compiled by Functional Analysis

Gene symbol	BKL description	GO Identifier
ACE	Angiotensin I converting enzyme 1; acts in GPCR pathway; regulates cell proliferation and blood pressure; induced inhibition prevents cardio vascular diseases; gene polymorphisms correlate with psoriasis; diabetes; and Alzheimer disease	GO:0002520
ANPEP	Alanyl aminopeptidase; involved in endothelial morphogenesis during angiogenesis; may play a role in cell proliferation; increased expression correlates with multiple sclerosis and non small cell lung cancer	GO:0006955
CD24	CD24 molecule; a signal transducer that acts in cell adhesion and induction of apoptosis; protein expression is upregulated in intrahepatic cholangiocarcinoma and several neoplasms; gene polymorphism is associated with multiple sclerosis	GO:0006955 GO:0002520
CD300C	CD300c molecule (CMRF-35 antigen); a cell surface membrane antigen and member of the immunoregulatory signaling family; plays a role in the cross-regulation of TNFalpha and IFNalpha secretion from plasmacytoid dendritic cells	GO:0006955
CD37	CD37 molecule; a putative transporter; mouse Cd37 is aberrantly expressed in B cell chronic lymphoproliferative disorders	GO:0006955
CD7	CD7 molecule; a signal transducer that acts in cell adhesion; immune response; and NK and T cell mediated cytotoxicity; altered expression correlates with lymphoproliferative disorders; pityriasis lichenoides; dermatitis; HIV infection; and sarcoidosis	GO:0006955 GO:0002520
CD8B	CD8b molecule; an MHC class I binding protein that plays a role in T cell differentiation; acts in cytokine and chemokine mediated signaling pathway; aberrant protein expression is associated with AIDS; renal cell carcinoma;	GO:0006955 GO:0002520

4.3.1 Expression, BKL manual curation

Each row of the table presents a matched organ or tissue. The columns contain (from left to right) the name of the organ or tissue linked to its Cytomer page, Gene symbols, Location (Tumors, Cell types or Organs, tissues, fluids), the number of input genes matching that group, the size of the matched group in Cytomer, the randomly expected number of hits and the P-value of the match result.

Figure 4.6 Output table of Expression analysis

Term	Gene symbol	Location	#Hits in group	Group size	#Hits expected	p-value
leukemia	ABCC1, ABCF2, ACE, ANPEP, AOAH, BAD, BCR, CASC3, CD24, CD7, ...	Tumors	43	1333	24	3.36846e-05
other tumors	ABCC1, ACE, ADD3, ADRM1, AKR1A1, ALDOB, ANPEP, ATF3, BAD, BECN1, ...	Tumors	114	5152	90	0.000206847
myeloid cells	ABCC1, ANPEP, AOAH, CD24, CD37, CD7, CTSG, CXCR3, DHCR24, ENO1, ...	Cell types	29	843	15	0.000276412
lymphoid cells	ABCC1, ACE, ANPEP, ARVCF, BAD, CASC3, CD24, CD300C, CD37, CD7, ...	Cell types	55	2031	36	0.00027954
sympathetic nervous system	HES1, SCN1B	Organs, tissues, fluids	2	2	1	0.000298033

4.3.2 Gene Ontology analysis (public and BKL curated)

Each row of the table presents a matched Gene Ontology term. The columns contain (from left to right) the GO identifier and a link to its description, Gene symbols, the GO term description, Ontology in general (Biological Process, Molecular Function or Cellular Component), the number of hits from the input set to the ontology group, the size of the group in the database, the number of randomly expected hits, and the P-value of the observation.

Figure 4.7 Output table of Gene Ontology classification

GO Identifier	Gene symbol	GO Term	Ontology	#Hits in group	Group size	#Hits expected	p-value
GO:0005515	ACE, ADD3, ADFP, AKR1A1, ALDOB, AP3B2, APRT, ARHGEF7, ARPC2, ARVCF, ...	protein binding	Molecular function	146	7814	96	3.50201e-12
GO:0008150	ABCC1, ABCF2, ACE, ADFP, ADRM1, AKR1A1, ALDOB, ANPEP, AOAH, AP3B2, ...	biological_process	Biological process	207	14410	177	9.73042e-10
GO:0005488	ABCC1, ABCF2, ACE, ADD3, ADFP, AKR1A1, ALDOB, ANPEP, AP3B2, APRT, ...	binding	Molecular function	188	12447	153	9.47439e-09
GO:0009987	ACE, ADRM1, AKR1A1, ALDOB, ANPEP, AOAH, AP3B2, APRT,	cellular process	Biological process	181	12076	148	2.14069e-07

4.3.3 Organ/tissue analysis output

Each row of the table presents a matched organ or tissue. The columns contain (from left to right) the name of the organ or tissue and a link to its Cytomer page, Gene symbols, the number of input genes matching that group, the size of the matched group in Cytomer, the randomly expected number of hits and the P-value of the match result.

Figure 4.8 Output table of Organ/Tissue expression analysis

	Organ/Tissue	Gene symbol	#Hits in group	Group size	#Hits expected	p-value ▲
☐	epidermis	ABCF2, ADFP, ADRM1, AKR1A1, APRT, ARPC2, ATF3, BAD, COX6A1, CSNK2B, ...	32	776	17	0.000256293
☐	tonsil	CHI3L2, COX6A1, COX7B, CSNK2B, DUT, ENO1, H3F3A, HCLS1, HLA-DPB1, NCF4, ...	17	314	7	0.00044935
☐	pancreatic islets	ADFP, C5orf13, CD24, CD37, CD7, DPEP1, EEF1B2, ENO1, FOLR2, GSTM1, ...	35	916	20	0.000540854
☐	choroid	ABCC1, ADFP, AKR1A1, ATP5G3, BCR, CHI3L2, CSNK2B, C5RP1, CTNND1, DBP, ...	32	854	19	0.00135381

4.3.4 BKL Disease analyses output

The table presents the matched disease in each row. The columns contain (from left to right) the MeSH ID and a link to its description page, Gene symbols, the descriptive name of the Disease, Biomarker associations type (correlation, causality or prevention), the number of input genes matching that group, the size of the matched group, the randomly expected number of hits, and the P-value of the match result.

Figure 4.9 Output table of BKL Disease analysis

	Disease	Gene symbol	Disease name	Biomarker associations	#Hits in group	Group size	#Hits expected	p-value ▼
☐	+ 70 rows on previous pages							
☐	D009303	HLA-E, TRAF3	Nasopharyngeal Neoplasms	Correlation	2	27	1	0.126786
☐	D023903	GNB3, IL1B	Coronary Restenosis	Correlation	2	27	1	0.126786
☐	D001172	HNRNPA2B1, ICAM2, IL1B, MMP14	Arthritis, Rheumatoid	Causality	4	84	2	0.126729

4.3.5 SwissProt analysis output

The table presents the the UniProt Knowledgebase keyword in each row. The columns are: the keyword ID and a link to its UniProt <http://www.uniprot.org/> description, Gene symbols, the descriptive name of the keyword, category, the number of input genes matching that keyword group, the size of the matched group, the randomly expected number of hits, and the P-value of the match result.

Figure 4.10 Output table of SwissProt analysis

	keyword ID	Gene symbol	keyword	category	#Hits in group	Group size	#Hits expected	p-value ▲
☐	KW0007	ADRM1, AKR1A1, ALDOB, APRT, EEF1D, ENO1, ETFB, FMO4, GTF2I, H3F3A, ...	Acetylation	PTM	35	906	12	1.32164e-09
☐	KW0480	MT1A, MT1H, MT1M, MT2A, MT3	Metal-thiolate cluster	Ligand	5	11	1	1.13271e-07
☐	KW0832	ADD3, ADRM1, DDB2, GTF2I, H3F3A, HIST1H1C, HIST1H2BC, HIST1H2BE, HIST1H2BK, HIST1H2BL, ...	Ubl conjugation	PTM	21	551	7	4.24426e-06
☐	KW0647	ADRM1, PSMB3, PSMB9, PSMC1, PSMD6, PSMD8	Proteasome	Cellular component	6	55	1	5.46774e-05
☐	KW0544	H3F3A, HIST1H2BC, HIST1H2BE, HIST1H2BK, HIST1H2BL, HIST1H2BN, LOC440926	Nucleosome core	Cellular component	7	86	2	8.76116e-05

4.3.6 Transcription Factor Classification analysis output

Each row of the table presents a matched class. The columns contain (from left to right) the transcription factor class identifier and a link to its TRANSFAC class browser, Gene symbols, the descriptive names

of the factor classes, the number of input genes matching that group, the size of the matched group in TRANSFAC, the randomly expected number of hits and the P-value of the match result.

Figure 4.11 Output table of Transcription Factor Classification analysis

	Factor class identifier	Gene symbol	Factor class description	#Hits in group	Group size	#Hits expected	p-value ▲
☐	1	ASCL1, ATF2, ATF3, ATF7, CEBPA, CREB1, CREM, FOSB, MAX, MYC, ...	Basic Domains	20	26	7	3.43991e-09
☐	3.5.2	ELK3, ELK4, ERG, ETS1, ETS2, SPI1	Ets-type	6	6	2	0.000238787
☐	1.1	ATF2, ATF3, ATF7, CEBPA, CREB1, CREM, FOSB	Leucine zipper factors (bZIP)	7	8	3	0.00035676
☐	1.3	MAX, MYC, MYCN, SREBF1, TFE3, USF2	Helix-loop-helix / leucine zipper factors (bHLH-ZI	6	7	2	0.00133692

4.3.7 TRANSPATH Molecule Classification analysis output

Each row in the table presents a matched class. The columns contain (from left to right) the Molecule class identifier (and a link to its TRANSPATH page), Gene symbols (input genes matching that group), Molecule class description, the number of input genes matching that group, the size of the matched group in TRANSPATH, the randomly expected number of hits and the P-value of the match result. Please note that classes are also molecule entries in TRANSPATH.

Figure 4.12 Output table of TRANSPATH Molecule Classification analysis

	Molecule class identifier	Gene symbol	Molecule class description	#Hits in group	Group size	#Hits expected	p-value ▲
☐	MO000037288	USF1, USF2	USF	2	2	1	0.00071026
☐	MO000082934	USF1, USF2	USF(h)	2	2	1	0.00071026
☐	MO000099581	ATF3, CREB1, DBP, MXD1, TCF3, USF1, USF2	Basic Domains	7	71	2	0.00255622
☐	MO000099590	MXD1, USF1, USF2	Helix-loop-helix / leucine zipper factors (bHLH-ZI	3	13	1	0.00437475

4.3.8 Whole subsets from the tree analysis output

Each row of the table presents a matched primary set node. The columns contain (from left to right) an identifier composed of the name of the project node (Geneset) and the name of the primary set node (Gene symbol), the number of input genes found in the set, the size of the primary set, the randomly expected number of hits, and the P-value of the match result.

Figure 4.13 Output table of Whole subsets analysis

	Geneset	Gene symbol	#Hits in group	Group size	#Hits expected	p-value ▲
☐	Down AnnMicroarray [233,104,431,158]	ABCC1, ACE, ALDOB, AOAH, AP3B2, BMP8A, BMP8B, C16orf35, CD300C, CPEB1, ...	44	233	3	2.49997e-44
☐	Human housekeeping genes [561,87,909,474]	AKR1A1, ARHGEF7, ARPC2, ATP5G3, BECN1, CASC3, COX6A1, CSNK2B, EEF1D, ENO1, ...	28	561	6	9.4405e-13
☐	Human housekeeping genes [562,87,909,468]	AKR1A1, ARHGEF7, ARPC2, ATP5G3, BECN1, CASC3, COX6A1, CSNK2B, EEF1D, ENO1, ...	28	562	6	9.85399e-13

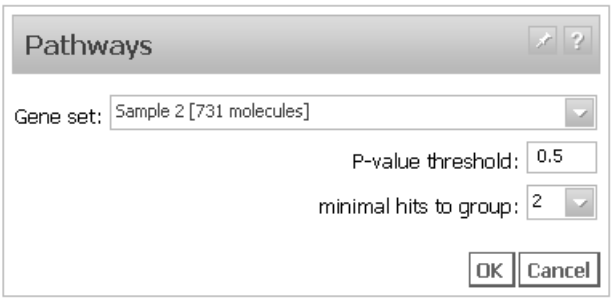
4.4 Functional Analysis with reaction pathways

The representation of molecules from reaction pathways can be used to investigate data sets. ExPlain provides you possibilities to choose either canonical TRANSPATH pathways or custom uploaded interaction sets.

4.4.1 Functional Analysis with TRANSPATH pathways

The "Canonical pathways mapping" option in menu "Analyze" or toolbar button can be used to investigate the representation of input molecules in annotated TRANSPATH pathways. As in the Functional classification, we can limit the output to pathways which received a minimal number of hits in the input set ("minimal hits to group" field), and to results with a minimal statistical significance ("P-value threshold" field). If there is an active data set in the project tree, it will be displayed in the "Gene set" field. You can choose another data set in this field (in this case the project tree has to be in multi-select mode). Clicking the OK button starts the process.

Figure 4.14 Pathways dialog



4.4.2 Pathways results

An example output is shown in Figure 4.15. Each row in the output table contains information about one pathway. Each pathway is referenced by name ("Pathway name" column) and id, where the "Pathway id" column links to the corresponding TRANSPATH entries. In the "Molecule name" column, molecules corresponding to certain pathway are listed. The "#Hits in group" column gives the number of input molecules found in a pathway, while values of the "Group size" column represent its total number of molecules. Results are sorted by P-value ("p-value" column). A "Visualization" column link will open a flash-based network visualizer (described in Section 7.3). The canonical pathway will be shown with the hits highlighted, as shown on Figure 4.16.

Figure 4.15 Pathways output

Pathway id	Molecule name	Pathway name	#Hits in group	Group size	#Hits expected	p-value ▲	Visualization
CH000000869	ABL-1a, ABL-1b, HMGB1, NF-YA, NF-YB, NF-YC-1, NF-YC-3, p73alpha, p73beta, Pin1, ...	p73 pathway	12	24	3	2.15813e-06	view
CH000000977	ABL-1a, ABL-1b, Delta40p53, p53-isoform1, p53beta, pre-ubiquitin, Ubc5B, UCRBP	Abl ---> p53	8	11	2	2.56188e-06	view
CH000003563	6-phosphofructokinase; liver type, Aldo3, aldolase-B, alpha-enolase, alpha-enolase-isoform2, fructose-bisphosphate aldolase A, GAPD, glucose-6-phosphate isomerase, phosphoglycerate kinase 1, PKM2, ...	glucose ---> pyruvate	11	22	3	5.90128e-06	view
CH000000974	Delta40p53, L11, L5, p53-isoform1, p53beta, pre-ubiquitin, Ubc5B, UCRBP	L11 ---> p53	8	12	2	6.95301e-06	view

4.4.3 Functional Analysis with user-defined interaction pathways

The "User defined pathways mapping" in menu "Analyze" links to analysis of the representation of input molecules in user defined interaction pathways, as described in Section 7.4. The dialog has the same parameters as the one mentioned above (Figure 4.14). The output table contains the column "Interaction pathways", where user interaction pathways are referenced by the pathway name. This column links to pathway data items in ExPlain. Columns "Molecule name", "#Hits in group", "Group size", "over(+)/under(-)representation" and "p-value" are identical to the ones from canonical TRANSPATH pathways output.

Figure 4.16 Visualization of the CH000000693 pathway

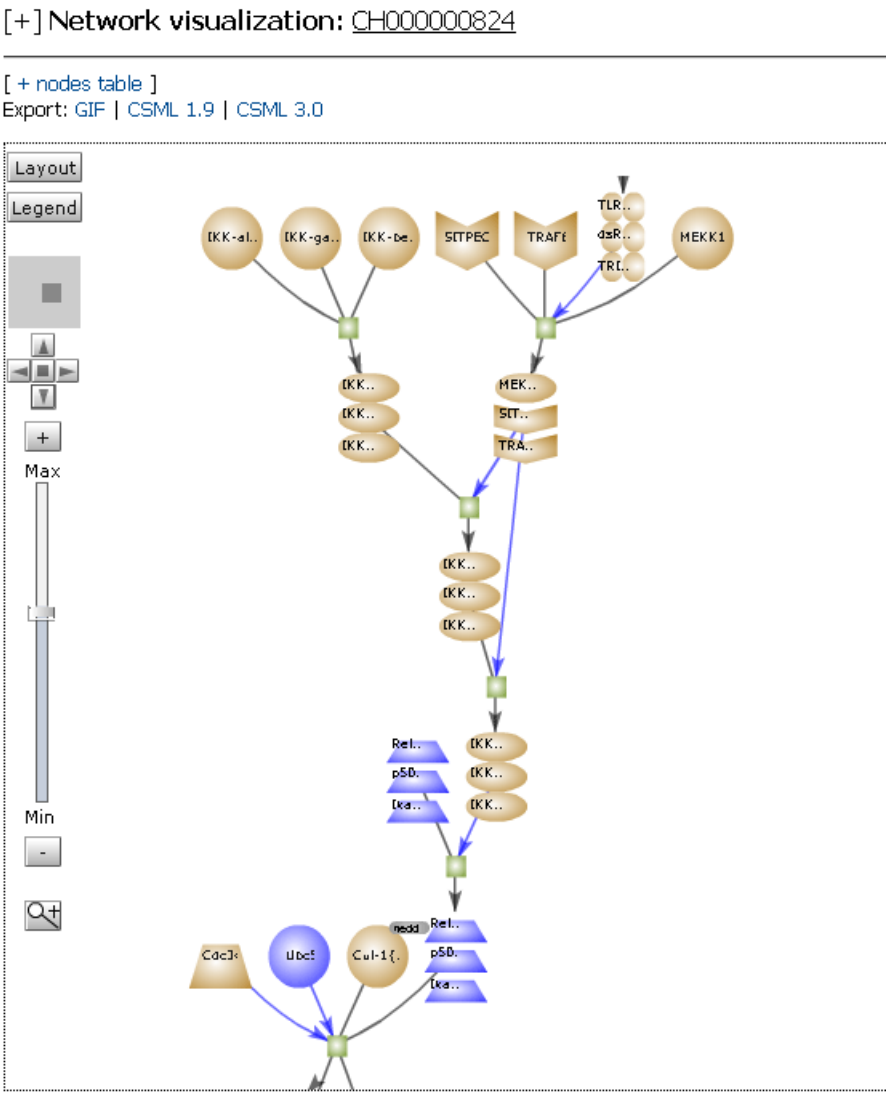


Figure 4.17 Interaction pathways output

[+] Pathways mapping: Interaction Pathways 1max 2min

Filter (filter bar): none (total 2 rows)

Rows per page: 10

Export: Plain text | XLS | RTF

Mark: All (2) | None | Invert

	Interaction pathway	Molecule name	#Hits in group	Group size	#Hits expected	p-value	Visualization
☐	Reactome_69615 [26]	ATM, Cdc25A, Cdk2, Delta40p53, mdm2-isoform1, p21Cip1, p27Kip1, p53-isoform1, p53beta, pre-ubiquitin, ...	12	14	3	8.26494e-10	view
☐	interactiontest [83]	Brm, CD38, grk4, HGF-isoform1, I-TAC, IL-12B, IL-13Ralpha2, MIP-3alpha, NAP1, p72, ...	18	42	7	2.66231e-07	view

4.5 Functional Analysis summary

ExPlain provides the possibility to compare classification results from several gene sets.

- Click the ”Functional classifications summary...” menu link in the ”Data” menu to launch the dialog window.

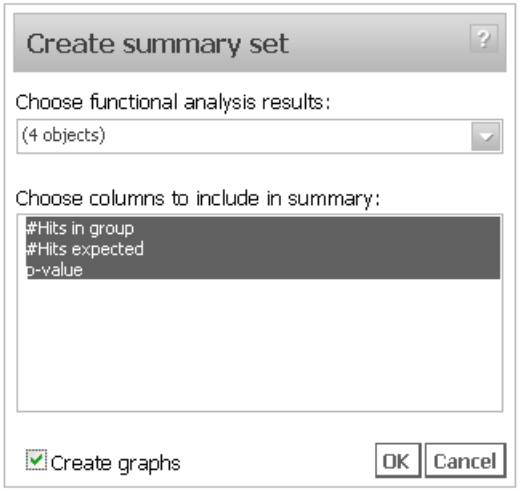
- Select analysis results.

- Select columns to include in the summary.

- Select checkbox for statistical graphs.

- Press ”OK” button.

Figure 4.18 FA summary dialog window



Each row of the output table is a group from FA analysis. The table contains common columns such as group size and group names, selected columns grouped by analyses results, and the ”sim/diff” column showing similarity between groups.

Figure 4.19 FA summary output table

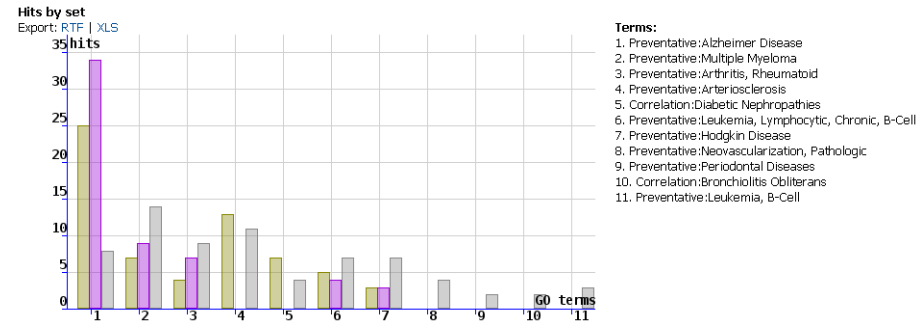
Result »					Proteome BKL Disease View 1max 2min on 'Sample 1' (Sample 1)			Proteome BKL Disease View 1max 2min on 'Human housekeeping genes' (Human housekeeping genes)			Proteome BKL Disease View 1max 2min on 'PXE' (PXE)			
	Disease	Disease name	Biomarker associations	Group size	#Hits in group	#Hits expected	p-value	#Hits in group	#Hits expected	p-value	#Hits in group	#Hits expected	p-value	sim/diff
☐	D000544	Alzheimer Disease	Preventative	148	25	11	4.36369e-05	34	24	0.0131628	8	4	0.00996028	sim
☐	D001161	Arteriosclerosis	Preventative	198	13	5	0.000552378	0	0	1	11	5	0.00208539	diff
☐	D001172	Arthritis, Rheumatoid	Preventative	84	4	2	0.126729	7	5	0.119798	9	2	4.30661e-05	diff
☐	D001989	Bronchiolitis Obliterans	Correlation	2	0	0	1	0	0	1	2	1	0.000408073	diff
☐	D003928	Diabetic Nephropathies	Correlation	52	7	2	0.000161005	0	0	1	4	2	0.0205804	diff
☐	D006689	Hodgkin Disease	Preventative	56	3	2	0.136984	3	3	0.528087	7	2	0.000118994	diff
☐	D009101	Multiple Myeloma	Preventative	160	7	4	0.074239	9	7	0.152966	14	4	3.33315e-06	diff
☐	D009389	Neovascularization, Pathologic	Preventative	15	0	0	1	0	0	1	4	1	0.000184174	diff

If the ”Create graphs” option was selected, the output also contains several graphical comparisons, as shown below.

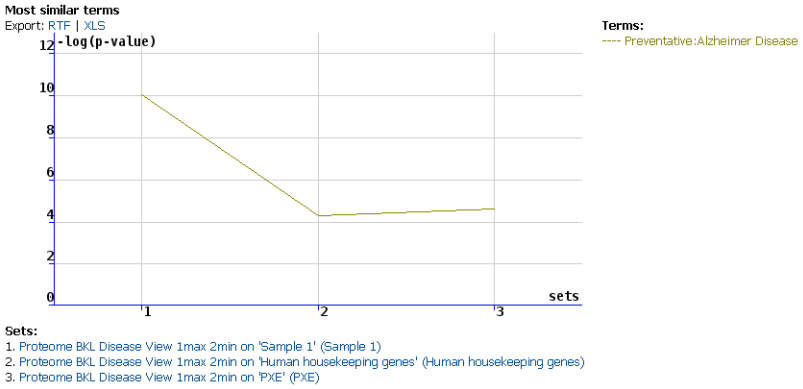
4.6 Functional Analysis algorithm

The Functional Analysis algorithm is designed to statistically analyze the representation of certain molecular classification groups in an input set. The input set is a list of genes. ExPlain can take any data set whose elements are mapped to the gene level automatically.

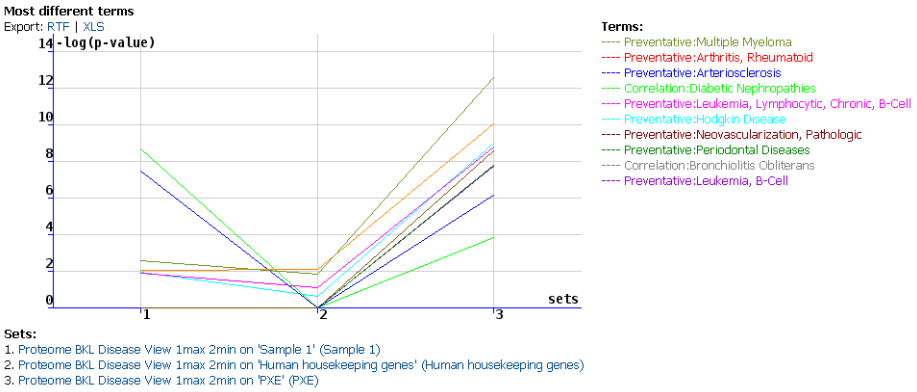
Figure 4.20 FA summary graphs



(a)


Most similar terms
Export: RTF | XLS
-log(p-value)
sets
Terms:
---- Preventative:Alzheimer Disease
Sets:
1. Proteome BKL Disease View 1max 2min on 'Sample 1' (Sample 1)
2. Proteome BKL Disease View 1max 2min on 'Human housekeeping genes' (Human housekeeping genes)
3. Proteome BKL Disease View 1max 2min on 'PXE' (PXE)

(b)


Most different terms
Export: RTF | XLS
-log(p-value)
sets
Terms:
---- Preventative:Multiple Myeloma
---- Preventative:Arthritis, Rheumatoid
---- Preventative:Arteriosclerosis
---- Correlation:Diabetic Nephropathies
---- Preventative:Leukemia, Lymphocytic, Chronic, B-Cell
---- Preventative:Hodgkin Disease
---- Preventative:Neovascularization, Pathologic
---- Preventative:Periodontal Diseases
---- Correlation:Bronchiolitis Obliterans
---- Preventative:Leukemia, B-Cell
Sets:
1. Proteome BKL Disease View 1max 2min on 'Sample 1' (Sample 1)
2. Proteome BKL Disease View 1max 2min on 'Human housekeeping genes' (Human housekeeping genes)
3. Proteome BKL Disease View 1max 2min on 'PXE' (PXE)

(c)

For a given functional category, such as Gene Ontology's Biological Process, the algorithm tests whether any group of this category is statistically over- or underrepresented in the data set.

For example, for the GO Biological Process "signal transduction" part of the Function category, the complete set consists of all genes which have a link to any group within this category. It also contains the maximal number of elements linked to the target group (like "signal transduction"). This is analyzed separately for each group within the category. Two P-values are computed for each set of hits to a category group, an overrepresentation (+) and an underrepresentation (-) P-value. They are derived from the hypergeometric distribution by Équation 4.6.1 and Équation 4.6.2. The Benjamini/Hochberg multiple testing correction [23] is applied to the list of obtained P-values.

In above equations

N the number of genes linked to the chosen category.

D the number of genes in the given group of the category.

quation 4.6.1 Overrepresentation of a Functional group

$$P(+) = \sum_{i=k}^{k_{max}} \frac{\binom{D}{i} \cdot \binom{N-D}{n-i}}{\binom{N}{n}}$$

$$k_{max} = \min(n, D)$$

quation 4.6.2 Underrepresentation of a Functional group

$$P(-) = \sum_{i=k_{min}}^{k} \frac{\binom{D}{i} \cdot \binom{N-D}{n-i}}{\binom{N}{n}}$$

$$k_{min} = \max(0, n-(N-D))$$

n the size of the input list.

k the number of genes that matched a gene in the group of the category.

The smaller the probability of observing the given input list (consisting of elements that are linked to the group or not), the stronger statistical significance of the assumption that the observation is not a random event. A probability threshold can be set in the P-value field of the ExPlain Functional Analysis interface.

In addition to groups found to be statistically over- or underrepresented according to the given P-value threshold, reported groups will be filtered by the minimal number of hits from the input set. This corresponds to the minimal hits to group parameter of the Functional Analysis dialog window.

4.7 Gene Set Enrichment Analysis

4.7.1 The GSEA interface

- Click the ”Gene Set Enrichment Analysis...” menu link in the ”Analysis” menu to launch the dialog window.

- Select a gene set that has at least one numerical column.

- Specify the numerical column to be used from the source gene set.

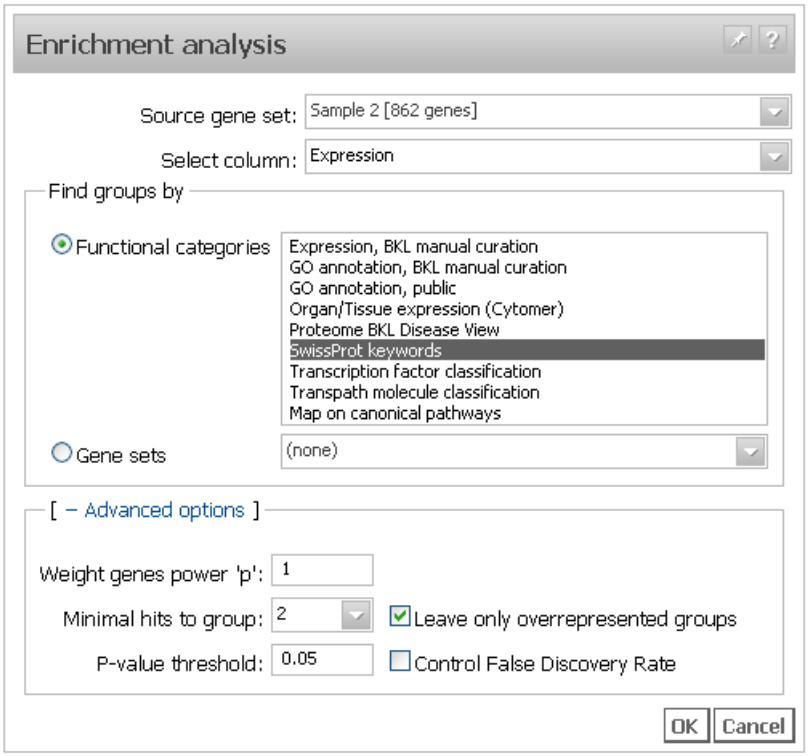
- Select ”Functional categories” groups to be searched for enrichment (see Section 4.1), molecules from canonical TRANSPATH pathways (see Section 4.4.1) or genes from desired ”Gene sets”.

- Press ”OK” button.

You can also specify more parameters in the ”Advanced options” block. The ”Weight genes power ’p’” parameter described in Équation 4.7.1 takes the values 0 or 1. To weight genes in the ”Source gene set” using the expression value, the p parameter should be set to 1, otherwise, set it to 0. Striving to keep this analysis in groups only essentially connected to the source gene set, you can control the ”Minimal hits to

group" and leave in the results "only overrepresented groups". If checked, overrepresented groups are searched using the FA algorithm from Section 4.6 preliminary to GSEA. To obtain statistically significant results, the "P-value threshold" should be set up close to zero. In the case of multiple groups testing (the most frequent case), you should use the "Control False Discovery Rate" option, so the "P-value threshold" will control the rate of false positives in the results.

Figure 4.21 GSEA user interface



4.7.2 GSEA example

Here we analyze a sample data set using an "Expression" column. We choose "SwissProt keywords" as a functional category. We set "Weight genes power" to 1 and left other parameters as they appear by default.

The enrichment result is pre-filtered by p-value, and you can clear the filter to see all lines. The figure below shows the enrichment output and detailed view for the "Transit peptide" keyword.

This result shows that genes associated with the "Transit peptide" function have low expression values in our experiment.

4.7.3 Enrichment Analysis theoretical basis

The Gene Set Enrichment Analysis (GSEA) algorithm is designed to detect the enrichment of input gene set with genes from certain molecular classification groups (or groups given by user). It is assumed that the input set contains genes differently expressed between two conditions (phenotypes, cell strains, before/after treatment etc). The input set D is a list of N genes with a value connected to each gene that reflects the difference in expression - it can be fold change, correlation with phenotype or other. Let's name these values "expression difference". The GSEA algorithm obtains a list of genes L by ranging the set D according to the expression difference (descending sort). L is a set of N genes {g_i} ordered by values {r_i}. For each group S of the given category, the algorithm accounts for whether hit genes of group S tend to be located at the top of the sorted list L, at the bottom or distributed randomly in L. While working the GSEA computes the Enrichment Score ES(S) that ranges in [-1,1] and p-value that ranges in [0,1]. ES is the maximum deviation from zero of the so called running sum Équation 4.7.1, [26].

As we can see, ES depends on the parameter p that ranges in [0,1]. If p=0 the ES resembles a Kolmogorov-Smirnov statistics [27]. If p>0 the algorithm takes into account the value of expression

Figure 4.22 Enrichment of a sample data set with "Transit peptide" genes

[+] **Enrichment results:** Enrich via group KWs p1 hits2

Filter (clear | filter bar): P-value < 0.05 (37 of 72 rows)

Export: Plain text | XLS | RTF

Mark: Page (10) | All (37) | None | Invert

keyword ID	keyword	category	Score	P-value ▲	Percent of hits	#Hits	Graphs
KW0372	Hormone	Molecular function	-0.790834	5.15e-08	1.2761	11	view
KW0539	Nucleus	Cellular component	0.183697	2.0565e-06	33.0626	285	view
KW0007	Acetylation	PTM	0.229931	2.8977e-06	16.3573	141	view
KW0809	Transit peptide	Domain	0.335462	0.000190063	4.40835	38	view
KW0963	Cytoplasm	Cellular component	0.166582	0.000243634	21.8097	188	view
KW0681	Retinal protein	Molecular function	-0.930151	0.000373296	0.348028	3	view
KW0600	Photoreceptor protein	Molecular function	-0.930151	0.000373296	0.348028	3	view
KW0157	Chromophore	Ligand	-0.930151	0.000373296	0.348028	3	view
KW0804	Transcription	Biological process	0.15751	0.0013919	18.4455	159	view
KW0238	DNA-binding	Ligand	0.158834	0.00146692	17.8654	154	view
+ 27 rows on next pages							
keyword ID	keyword	category	Score	P-value ▲	Percent of hits	#Hits	Graphs

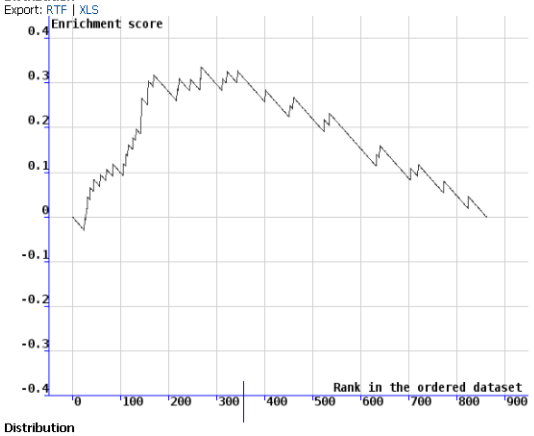
(a)

[+] **Enrichment results:** Enrich via group KWs p0 hits2

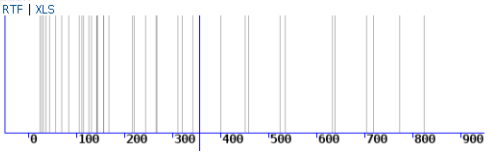
[<< Back to group list]

keyword ID: KW0809
keyword: Transit peptide
category: Domain
Score: 0.335462
P-value: 0.000190063
Percent of hits: 4.40835
#Hits: 38
Hits: ATP5A1, ATP5D, ATP5G1, ATP5G3, ATP5O, C21orf33, COX4I1, COX5A, COX5B, COX6A1, COX6A2, COX7A2, COX7A2L, COX7C, COX8A, CYC1, GOT2, GPX4, HADHA, HADHB, IDH3B, KIAA0141, MRPL9, MRPS12, NDUFC1, NDUFV1, NDUFV2, SDHA, SLC25A1, SLC25A3, STARD7, TIMM44, TRAP1, TSFM, TUFM, UQCRC1, UQCRFS1, UQCRH

Distribution
Export: RTF | XLS



Distribution
Export: RTF | XLS



(b)

difference of each gene. P-value assesses the probability to get a similar or better result by chance and should be restricted with a cut-off. The P-value calculation has 3 cases. p=0 and single S testing. The P-value is calculated precisely using a special dynamic programming algorithm that counts all possible cases when the absolute value of ES is higher than the absolute value of the given ES if list L has another order, and divides the number of such cases by the number of all possible variants of order in L.

p>0 and single S testing. Expression difference values {r_i} are reassigned to genes {g_i} randomly in

quation 4.7.1 Enrichment Score

$ES(S)$ – is the maximum deviation from zero across i of the running sum

$$ES(S,i) = P_{hit}(S,i) - P_{miss}(S,i)$$

$$P_{hit}(S,i) = \sum_{\substack{g_j \in S \\ j \le i}} \frac{|r_j|^p}{N_R}, \quad P_{miss}(S,i) = \sum_{\substack{g_j \notin S \\ j \le i}} \frac{1}{(N - N_H)}, \quad where$$

$$N_R = \sum_{g_j \in S} |r_j|^p, \quad N_H - number of\ hits\ S\ in\ L$$

D and ES is recalculated. A permutation test performs this procedure 1000 times and builds a histogram of the corresponding enrichment scores - ES_null distribution histogram. The P-value is evaluated by using the positive or negative tail of the distribution corresponding to the sign of observed ES.

Multiple S testing. The most common case - a permutation test with adjustment for variation in multiple S sets size and FDR control of false positives. For each set S, ES_null(S) distribution histogram is built as described above. Each ES(S) is adjusted by dividing it by the positive or abs negative mean of scores obtained in permutations depending on the sign of observed ES(S) (positive and negative scores take part in positive and negative mean calculation separately). Using the internally built ES_null histograms p-values are calculated and by default adjusted to control False Discovery Rate (FDR) [23]. If ES is significantly greater than 0, hits of gene set S tend to be located at the top of sorted L list, if ES is significantly lower than 0 - hits of set S tend to be located at the bottom of L.

Chapter 5

Transcription factor site search

The site analysis module allows you to find putative transcription factor binding sites (TFBS) in the promoters of a gene set, interval set, or user-loaded sequences. Binding site predictions are obtained by running MATCH, Patch, or P-Match, which use collections of known TFBS and positional weight matrices to identify the most probable binding sites. Seeder is used to find de-novo motifs overrepeperesented in the analysed set. Promoters are automatically extracted by ExPlain from the TRANSPro database according to specified ranges upstream and downstream from the virtual transcription start site (TSS), and considering the extraction rules applied to genes with multiple promoter entries. In these analyses, two sets of promoters, a query (positive) and a control (negative) set, will be analyzed to enable you to identify factors with overrepresented sites or motifs in the query set.

5.1 Searching for sites in promoters

In order to find which transcription factors might control the genes in the dataset, ExPlain will find the TFBSs in this set and compare the frequencies of these sites with those in a control or background gene set (non-changed genes). This will produce a collection of Position Weight Matrices (PWM) whose described sites are overrepresented in the promoter set under study.

NOTE

You are not confined to do a differential analysis and require a control set. The data set may as well be a single group of genes of interest or a single sequence, and the report will then contain the absolute frequencies of the binding sites found.

5.1.1 How to run Match

- Click the ”Match” menu link in the ”Analysis” menu or press to launch the dialog window.

- Select analysed set (upregulated genes for example) as Yes set.

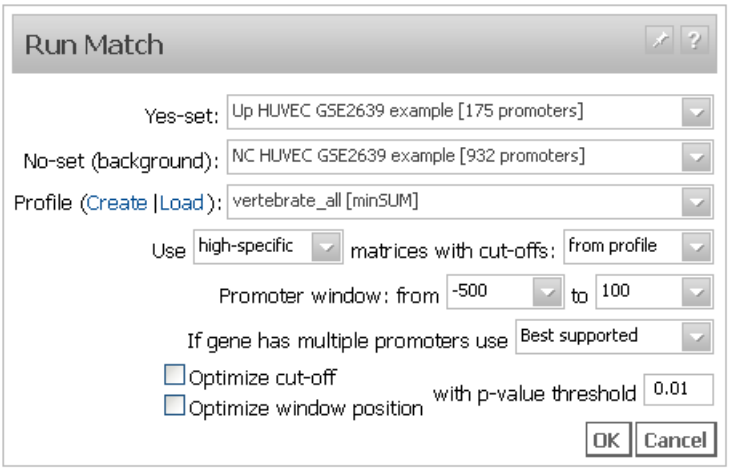
- Select background set (not changed genes for example) as a No set.

- Press ”OK” button.

- To extract a list of transcription factors select the matrices with a good yes/no ratio and low p-value and press icon.

5.1.2 Sites search dialog window

Figure 5.1 The Match dialog window



Click the "Match" menu link in the "Analysis" menu to launch the dialog window. The dialog provides fields to select a profile of TRANSFAC matrices for the search, the promoter window around the TSS, and to specify a rule for promoter selection, if several are available for an individual gene. Select the main set you want to analyze in the "Yes-set" field. You can choose a gene set, an interval set or a set of loaded sequences. If you are standing on an appropriate data set tree node when the sites search analysis dialog is launched, the "Yes-set" field is set by default to the current gene set. We recommend you specify a "No-set" to use as background set for calculations, but it's not strictly required. By default this field is set to "none" (or to the last gene set used as background before, if any).

NOTE

Promoters present in both (Yes and No) sets will be excluded from the background.

The profile list contains all profiles from the whole project tree. To create a new profile or load an existing one from a file, use the Create | Load link. You can select score thresholds of the PWMs from the profile to use in your analysis. Cut-offs from the original profile, minimized false negatives (minFN), sum of FP and FN (minSUM) and minimized false positives (minFP) are available for the selection. Note that the original profile will not be changed. To find an individual cut-off for each PWM that provides the best frequency ratio between the analysed and background set, use the "Optimize cut-off" option (see Section 5.1.4).

The maximal size of a TRANSPro promoter sequence available is 11000 nucleotides, ranging from position 10000 upstream to 1000 downstream of the TSS defined by TRANSPro. This means that the maximal value in "from:" is -10000 and the maximal value in to: is 1000. When you upload your own seqeunces, the maximal size range allowed is 200000 nucleotides. Within this range, you can set the TSS arbitrarily at any position. To find an individual promoter window for each PWM providing the best separation of the query and background sets, use the "Optimize window position" option (see Section 5.1.4).

By default, ExPlain searches the promoter sequences of every gene provided by TRANSPro, where each promoter corresponds to the best supported virtual TSS. You can otherwise change restriction to use 5′ most, 3′ most, or all promoters.

After setting up the required parameters, press the OK button to start the analysis. To leave the dialog window open after starting a process and schedule a second one, press the pin button, modify the parameters as needed and then send the analysis again.

NOTE

Do not use optimization options when you plan to run CMA on the results of this search.

5.1.3 The Match output

In this result we show the MATCH output generated from the analysis of promoter sets extracted from the "HUVEC GSE2639 example" gene set. We extracted the top 100 up-regulated genes as query and 500 genes with FC˜1 as a background set. All parameters were set as described in Section 8.2.2. Results are accessible at the MATCH node added to the project tree below the node of the query set ("Yes set").

In the output table "Yes" refers to the main set and "No" refers to the background set. Each row contains information about the performance of one matrix of the input PWM profile. Matrix names link to corresponding matrix entries in the BKL database. The average number of putative binding sites per 1000 bp is given for the query and the background sets in "Yes" and "No" columns respectively. Additionally, "Yes" and "No" values are visualized in the "Graphs" column at the right, where the red bar depicts the abundance of the PWM motif in promoters of the positive set and the blue bar displays its abundance in the control set. The ratio of the two values is provided in the "Yes/No" column, where a number greater than one indicates overrepresentation of the motif in the query set. Significance of the representation value is measured by the P-value derived from a binomial distribution. "Matched promoters p-value" assesses the statistical significance of the number of promoters in the query set that have at least one predicted site compared to that of promoters in the control set.

Figure 5.2 Site search analysis output table

	Matrix name	Yes (sites/1000bp)	No (sites/1000bp)	Yes/No	P-value ▲	Graphs	Matched promoters p-value
	V$NFKAPPAB65_01	1.6833	0.5464	3.0805	7.2224e-17		3.2912e-08
	V$NFKB_Q6	3.0167	1.4686	2.0541	1.0084e-14		0.0020
	V$RELBP52_01	1.1333	0.3279	3.4567	1.5155e-13		1.2663e-08
	V$NFKB_Q6_01	0.9167	0.2391	3.8343	1.7253e-12		9.5913e-09
	V$NFKAPPAB_01	2.5333	1.2534	2.0211	3.2929e-12		1.7481e-04
	V$CREL_01	1.5500	0.6352	2.4400	2.7677e-11		1.2375e-06
	V$P50RELAP65_Q5_01	1.5500	0.6626	2.3394	1.5430e-10		5.4541e-05
	V$NFKB_C	1.8833	0.8948	2.1047	2.7717e-10		1.5448e-04
	V$P50P50_Q3	1.2333	0.5567	2.2155	5.9381e-08		6.7420e-05
	V$NFKAPPAB50_01	4.6500	3.3572	1.3851	1.8539e-06		0.0162
	V$IRF7_01	0.9500	0.4508	2.1073	5.9623e-06		0.0031
	V$POU1F1_Q6	1.5833	0.9324	1.6982	1.3464e-05		0.0520

Some additional information about the analysis results is available in hidden columns. You can read about hiding or showing columns in Section 1.5.5. "Matched promoters in Yes" and "Matched promoters in No" shows the fraction of promoters where at least one site of the matrix was found in the main and background sets respectively.

An example output table of an analysis done without a control set is shown below Figure 5.3. Each row contains information about the performance of one matrix of the input PWM profile. Matrix names link to corresponding matrix entries in the BKL database. The average number of matches per 1000bp is given for in the "Sites density" column and is visualized in the "Graphs" column.

Use the "Site search result" menu to generate a graphical visualization of the results and various subsets of information from the output table for the selected rows. The "Gene set" menu link creates a gene

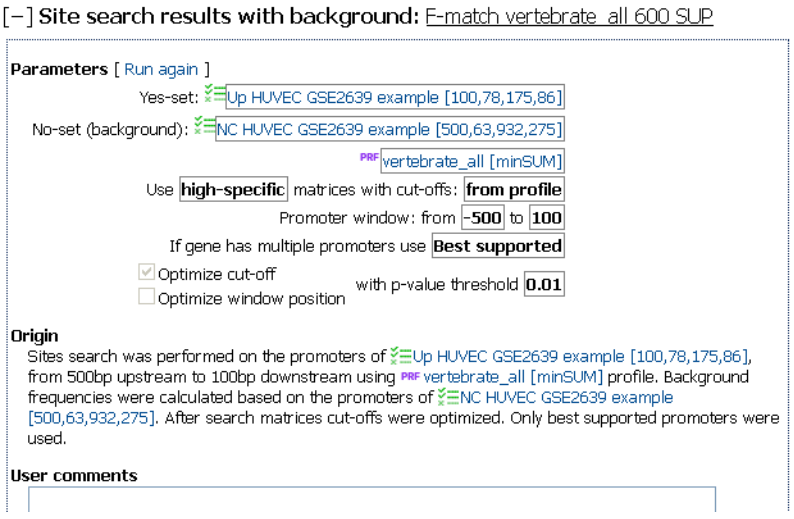
Figure 5.3 Match output for query set only

	Matrix name	Sites density (per 1000bp) ▼	Graphs
☐	V$ZF5_01	40.6167	
☐	V$WT1_Q6	30.6667	
☐	V$ZF5_B	21.8000	
☐	V$LRF_Q2	20.7333	
☐	V$AP2ALPHA_03	19.6167	
☐	V$ZIC1_01	17.7667	

set of selected matrices, the "Homologous matrices" extends the gene set by adding all the transcription factors linked to homologous matrices. Note that in both actions columns from site search results will be added to the created set. The "Site map" menu link opens a new window with a report as described in Section 5.2 for all selected matrices together. The "Save matrices" menu link creates detailed reports for all of the selected matrices. The "Sites table" displays the formatted MATCH output (see Figure 5.24) for selected matrices and for all promoters used in the analysis. The "Profile" option creates a PWM profile with the cut-offs from the result. Matrices from MATCH output can be exported to BKL as search result with "BKL search result" (more about export you can read in Section 3.5).

Parameters of the search can be recalled by expanding additional information (see Section 1.5.2). There are also some descriptions of the analysis and a field for comments.

Figure 5.4 Site search analysis additional information


[−] Site search results with background: F-match vertebrate_all 600 SUP
Parameters [Run again]
Yes-set: Up HUVEC GSE2639 example [100,78,175,86]
No-set (background): NC HUVEC GSE2639 example [500,63,932,275]
PRF vertebrate_all [minSUM]
Use high-specific matrices with cut-offs: from profile
Promoter window: from -500 to 100
If gene has multiple promoters use Best supported
Optimize cut-off
Optimize window position
with p-value threshold 0.01
Origin
Sites search was performed on the promoters of Up HUVEC GSE2639 example [100,78,175,86], from 500bp upstream to 100bp downstream using PRF vertebrate_all [minSUM] profile. Background frequencies were calculated based on the promoters of NC HUVEC GSE2639 example [500,63,932,275]. After search matrices cut-offs were optimized. Only best supported promoters were used.
User comments

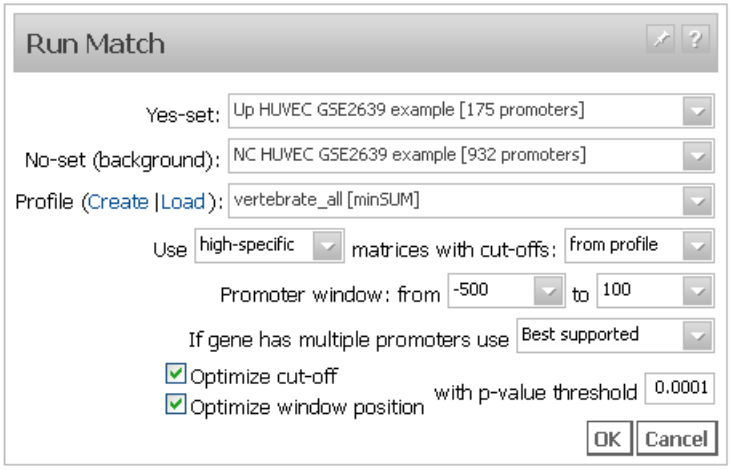
When the MATCH output is large and the result is not activated (as the in the case of a run within the CMA analysis) for more than 3 days, it will be compressed by ExPlain in order to save disk space. In this case some menu features will be disabled, but it is still possible to view the main Site search output and create a set or profile from the significant matrices. To restore full functionality, use the "Restore full result options" menu option of the Site search specific menu or the "Restore" link on the Site search item page. The waiting time period before a MATCH result is compressed can be adjusted by the server administrator.

5.1.4 Optimization mode

The "Optimize cut-off" option turns on the profile optimization mode. In this mode matrices which do not differ much in density between positive and negative set will be removed from the result. Cut-offs of the remaining matrices will be optimized to maximize the difference between the query and control sets. By adjusting the "p-value threshold" option you can vary the severity of this effect. Lower values of "p-value threshold" correspond to strict filtering, so you will get only a few very significant matrices. Higher values correspond to loose filtering; in this case you will get more entries in the result. The

"Optimize window position" option turns on an algorithm to find out the window that maximizes the difference between query and control sets. The initial window size is set to 300 bp and is increased by 100 bp on each iteration. The window slides from downstream to upstream position with a step of 100 bp. The p-value is calculated for all available combinations of window size and position, and the best one is taken as the result.

Figure 5.5 Sites search dialog window with optimization turned on



If the "Optimize window position" option was checked, the output table contains optimized upstream and downstream positions in "From" and "To" columns respectively.

Figure 5.6 Optimized site search result table

Filter (filter bar): none (total 41 rows)

Rows per page: 10

Export: Plain text | XLS | RTF | Raw output

1 2 3 4 5

Mark: Page (10) | All (41) | None | Invert

	Matrix name	Yes (sites/1000bp)	No (sites/1000bp)	Yes/No	P-value ▲	Graphs	Matched promoters p-value	From	To
	V$CREL_01	0.3167	0.0034	92.7200	4.0395e-14		2.4823e-12	-200	100
	V$NFKB_C	0.2833	0.0034	82.9600	1.2584e-12		3.8639e-13	-300	100
	V$NFKB_Q6_01	0.2833	0.0034	82.9600	1.2584e-12		3.8639e-13	-300	100
	V$NFKAPPAB_01	0.2833	0.0034	82.9600	1.2584e-12		3.8639e-13	-300	100
	V$NFKAPPAB65_01	0.3000	0.0068	43.9200	1.8941e-12		5.1255e-13	-300	100
	V$RELBP52_01	0.4500	0.0546	8.2350	2.5611e-11		4.1206e-11	-300	0
	V$P50RELAP65_Q5_01	0.2167	0.0068	31.7200	7.4143e-09		3.9078e-09	-200	100
	V$NFKB_Q6	0.1833	0.0034	53.6800	3.4868e-08		7.5009e-07	-300	0
	V$IRF_Q6	0.2000	0.0102	19.5200	1.5968e-07		2.4842e-06	-200	100
	V$STAT1_01	0.6167	0.2015	3.0603	4.3130e-07		3.1720e-05	-400	0
	+ 31 rows on next pages								
	Matrix name	Yes (sites/1000bp)	No (sites/1000bp)	Yes/No	P-value ▲	Graphs	Matched promoters p-value	From	To

The result of an optimized site search contains optimized PWMs cut-off values within the hidden columns. Use the "Profile" menu option to create a PWM profile with optimized cut-off values.

5.1.5 P-Match: combine pattern and matrix search

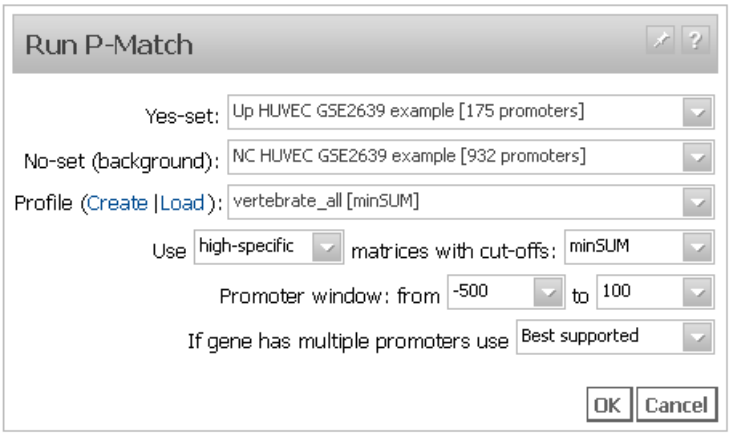
As an alternative to the MATCH algorithm, you can use P-Match, which combines pattern matching and weight matrix approaches (see Section 5.4.6). In comparison with the MATCH approach, P-Match generally provides superior recognition accuracy in the area of low false negative errors (high sensitivity), while in the area of low FP errors MATCH performs better in site recognition accuracy.

To launch the P-Match dialog window click the "P-Match" menu link in "Analyse" menu. As is the case for the MATCH algorithm, the dialog provides fields to select a profile of TRANSFAC matrices

for the search, the cut-off level for matrices, the promoter window and to specify a rule for promoter selection if several are available for an individual gene. After setting up required parameters, press the OK button to start the analysis.

> NOTE
>
> Only matrices that have a defined collection of sites will be used for the search, so the number of matrices in the result can be less than in the profile.

Figure 5.7 P-Match dialog window



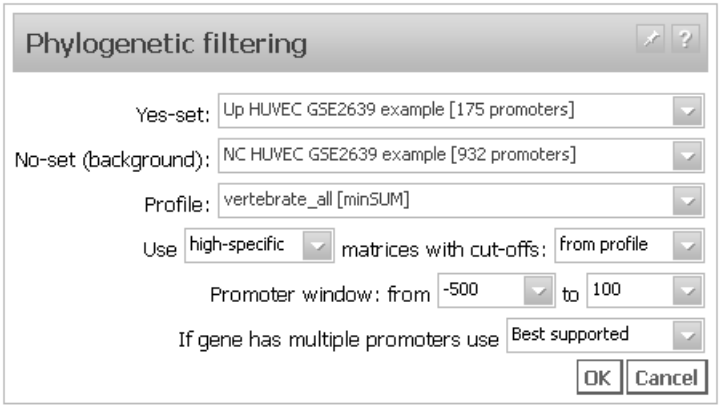
The Output table and further analyses options for the P-Match results are the same as for the MATCH (see Section 5.1.3).

5.1.6 Searching for sites in promoters using phylogenetic filtering

MATCH combined with footprint method uses homology information between different species (Human, Mouse and Rat in our case) to reveal binding sites that are present in homologous promoters of all organisms.

To run the analysis, press the "Phylogenetic filtering" menu link in the main menu "Analyze" . As you see below, the dialog options are almost the same as in Section 5.1.2. You should set up a query gene set, control set, promoter window and promoter type.

Figure 5.8 Site search with footprint dialog window



For each promoter of the main gene set the algorithm leaves only those TF binding sites that were found in homologous promoters of other species. First the MATCH algorithm searches for TF binding

sites in all sets of promoters. Then the search results are processed to get the intersection in terms of matrices. Thus, there can be fewer matches for matrices in the output table in comparison with the regular sites search analysis result (see Figure 5.2).

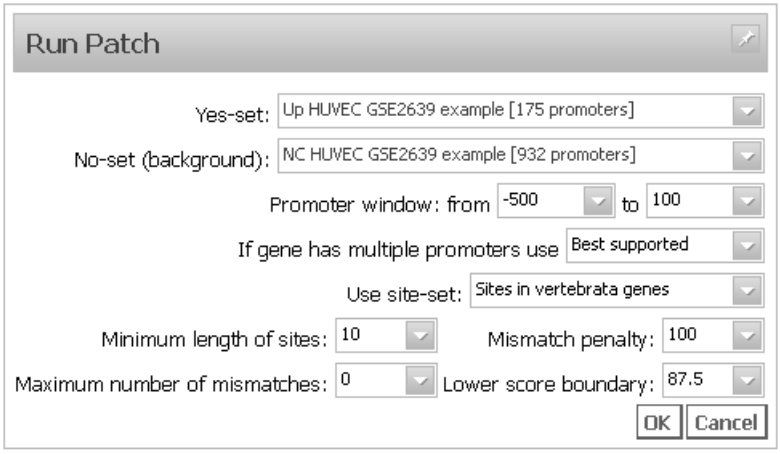
Figure 5.9 Output table of footprint site search

	Matrix name	Yes (sites/1000bp)	No (sites/1000bp)	Yes/No	P-value ▲	Graphs	Matched promoters p-value
	V$NFKB_Q6	1.3333	0.3242	4.1129	1.4104e-17		1.4242e-04
	V$NFKAPPAB65_01	0.7018	0.1106	6.3448	1.1011e-13		2.5896e-06
	V$NFKAPPAB_01	1.0351	0.2593	3.9912	1.2930e-13		9.1888e-04
	V$NFKB_Q6_01	0.3684	0.0114	32.2000	2.2433e-13		6.0070e-09
	V$CREL_01	0.5965	0.0763	7.8200	2.6121e-13		4.2066e-07
	V$NFKB_C	0.6667	0.1068	6.2429	6.0654e-13		6.1471e-06
	V$RELBP52_01	0.4211	0.0267	15.7714	7.8041e-13		3.4842e-07
	V$P50RELAP65_Q5_01	0.5439	0.0992	5.4846	5.6912e-10		2.3197e-05
	V$IRF7_01	0.3684	0.0496	7.4308	1.5983e-08		0.0015
	V$P50P50_Q3	0.3860	0.0572	6.7467	1.9728e-08		0.0011
	V$STAT5A_02	2.5965	1.5866	1.6365	4.8714e-07		0.0318
	V$NFKAPPAB50_01	2.1754	1.3272	1.6391	3.5936e-06		0.0596
	V$SREBP2_Q6	6.8772	8.6346	0.7965	1.1210e-05		0.0370
	V$TAL1BETAE47_01	1.3860	2.1739	0.6375	4.8541e-05		0.0161
	V$KID3_01	3.5789	4.7635	0.7513	5.3801e-05		0.0451
	V$E2F1DP2_01	0.0175	0.2288	0.0767	8.7681e-05		0.0051
	V$CACCCBINDINGFACTOR_Q6	1.8421	1.1899	1.5481	1.1569e-04		0.4349

5.1.7 Pattern based search

Patch searches for TF binding sites by comparison of the analysed sequences with a collection of known sites. To launch the Patch dialog window click the "Patch" menu link in the "Analyse" menu. Like for the MATCH (see Section 5.1.2) you can select datasets to analyse and set promoter parameters. . After setting up the required parameters, press the OK button to start the analysis.

Figure 5.10 P-Match dialog window



In the output table, "Yes" refers to the main set and "No" refers to the background set. Each row contains information about the performance of one site. Site names link to corresponding site entries in BKL. Binding factors for the sites are given in the "Binding factors" column. The average number of matched sites per 1000bp is given for the main and the background sets in "Yes" and "No" columns respectively. Additionally, "Yes" and "No" values are visualized in the "Graphs" column, where the red bar depicts the abundance of the PWM motif in promoters of the query set and the blue bar displays its

abundance in the control set. The ratio of the two values is provided in the "Yes/No" column, where a number greater than one indicates overrepresentation of the motif in the analysed set. Significance of the representation value is measured by the P-value derived from a binomial distribution. "Matched promoters p-value" assess the statistical significance that each "Yes" promoter presents at least one site when compared with "No" promoters.

Figure 5.11 Patch analysis output table

Filter (filter bar): none (total 1082 rows)

Rows per page: 10

Export: Plain text | XLS | RTF | Raw output

« 1 2 3 4 5 6 7 8 »

Mark: Page (10) | All (1082) | None | Invert

	Binding factors	Site name	Yes (sites/1000bp)	No (sites/1000bp)	Yes/No	P-value ▲	Graphs	Matched promoters p-value
☐ (*)	HMGIY	HS$IL2RA_12	0.0000	0.2152	0.0000	7.9374e-06		0.1526
☐ (*)	NF-kappaB1-p50	HS$NFKB1_02	0.1167	0.0034	34.1600	2.8020e-05		2.3672e-05
☐ (*)	NF-kappaB	HS$CXCL1_01	0.1167	0.0034	34.1600	2.8020e-05		2.3672e-05
☐ (*)	(RelA-p65)2, c-Rel:RelA-p65, NF-kappaB, NF-kappaB1-p50:c-Rel, NF-kappaB1-p50:RelA-p65	HS$CCL2_06	0.1167	0.0034	34.1600	2.8020e-05		2.3672e-05
☐ (*)	NF-kappaB	HS$GRO3_01	0.1167	0.0034	34.1600	2.8020e-05		2.3672e-05
☐ (*)	RelA-p65	MOUSE$CCL2_02	0.1167	0.0034	34.1600	2.8020e-05		2.3672e-05
☐ (*)	NF-kappaB	HS$CXCL2_01	0.1167	0.0034	34.1600	2.8020e-05		2.3672e-05
☐ (*)	NF-kappaB, RelA-p65:RelA-p65	HS$CD40_06	0.1167	0.0034	34.1600	2.8020e-05		2.3672e-05
☐ (*)	NF-kappaB, NF-kappaB1-p50	RAT$SAA1_06	0.1333	0.0102	13.0133	7.0704e-05		5.7620e-05
☐ (*)	sp3	RAT$GRIN1_07	0.1333	0.0102	13.0133	7.0704e-05		5.7620e-05
☐	+ 1072 rows on next pages							
	Binding factors	Site name	Yes (sites/1000bp)	No (sites/1000bp)	Yes/No	P-value ▲	Graphs	Matched promoters p-value

Use the "Site search result" menu to show a graphical match visualization and various subsets of information from the output table for selected rows. The "Site map" menu link opens a new window with a report as described in Section 5.2 for all selected sites together. The "Save matrices" link creates detailed reports for each of the selected sites. Columns with the number of sites found for each gene will be automatically added to the "Yes" set. The "Sites table" displays the formatted MATCH output (see Figure 5.24) for selected matrices and for all promoters used in the analysis.

5.1.8 De-novo motifs identification

Use the Seeder algorithm to discover new motifs overrepresented in your dataset. To launch the Seeder dialog window, click the "Seeder" link in the "Analyse" menu. Like in the case of MATCH (see Section 5.1.2) you can select datasets to analyse and promoter parameters. Length of DNA sequences to analyse should not be larger than 50000 bp. A suitable input set would be up to 100 promoter sequences each of 500-600 bp. ExPlain will automatically constrain the promoter window when the number of selected promoters is high. Typically 50-60 sequences will show well conserved motifs. If the expression of the corresponding genes is known, a selection of top 50 upregulated genes will work best. There are no size restrictions regarding the background promoter set, yet a rich background provides good quality and low false discovery rate. After setting up the required parameters, press the OK button to start the analysis.

The resulting table contains a selected number of motifs with the names of the corresponding matrix. Marices are saved as separate tree nodes below the Seeder result. When the cut-off calculation procedure is finished, newly created matrices can be included in a profile and used for MATCH search.

5.1.9 Filtering site search results using intervals

The dialog window "Filter Match results" appears after clicking on the link "Filter sites using intervals" in the "Analyze" menu. This dialog provides fields for selection of a sites search result which you want to filter. All the previously uploaded intervals (see Section 9.1) are available through the "Interval set to use" dropdown box. After filtering, you will get just binding sites that are found within an interval. You can choose the option to leave binding sites that are located in an interval linked to the transcription factor binding this site. It is also possible to extend the intervals using the "Expand interval by" field.

Filtering of the results obtained in the example mentioned above with a predefined interval set.

Figure 5.12 Seeder dialog window

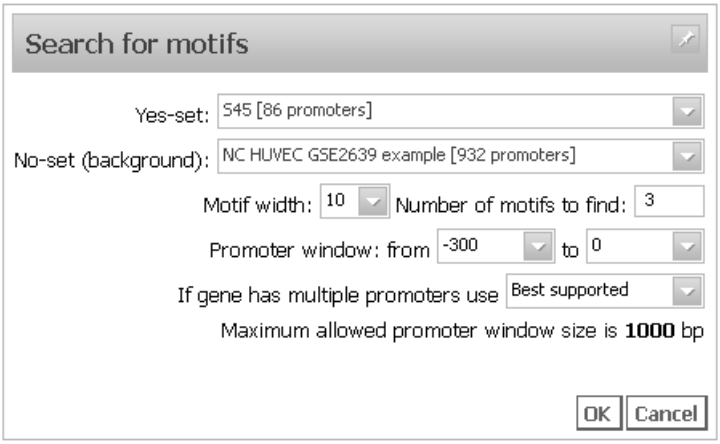
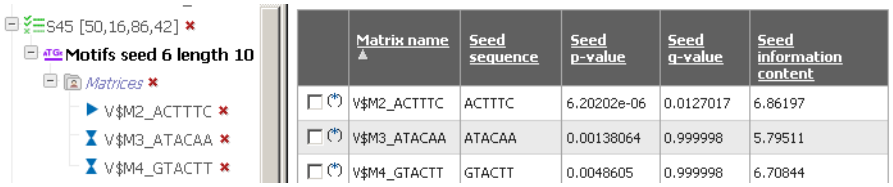


Figure 5.13 Seeder output table



	Matrix name	Seed sequence	Seed p-value	Seed q-value	Seed information content
	V$M2_ACTTTC	ACTTTC	6.20202e-06	0.0127017	6.86197
	V$M3_ATACAA	ATACAA	0.00138064	0.999998	5.79511
	V$M4_GTACTT	GTACTT	0.0048605	0.999998	6.70844

Figure 5.14 Matrix created with Seeder .

Accession ID: **X15002**
Matrix quality:
Window size: **10**
Binding factors: **none** (change)

	1	2	3	4	5	6	7	8	9	10
A	12	10	42	3	1	0	0	2	12	13
C	2	8	0	30	2	1	7	47	21	12
G	25	31	5	2	1	0	0	1	1	17
T	11	1	3	15	46	49	43	0	16	8
Consensus	G	G	A	C	T	T	T	C	N	N
Picture										

Figure 5.15 Sites search filtering dialog window

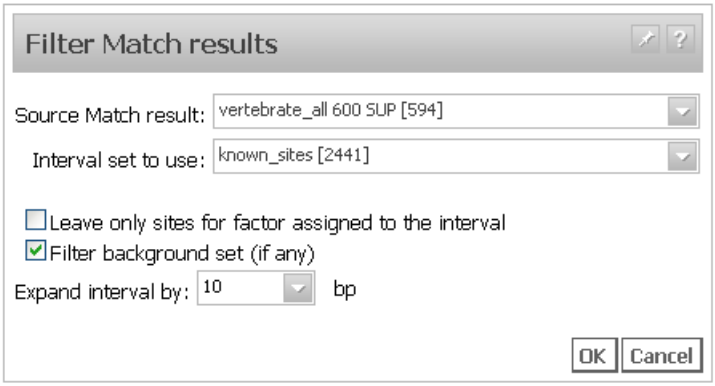


Figure 5.16 Sites search filtering results

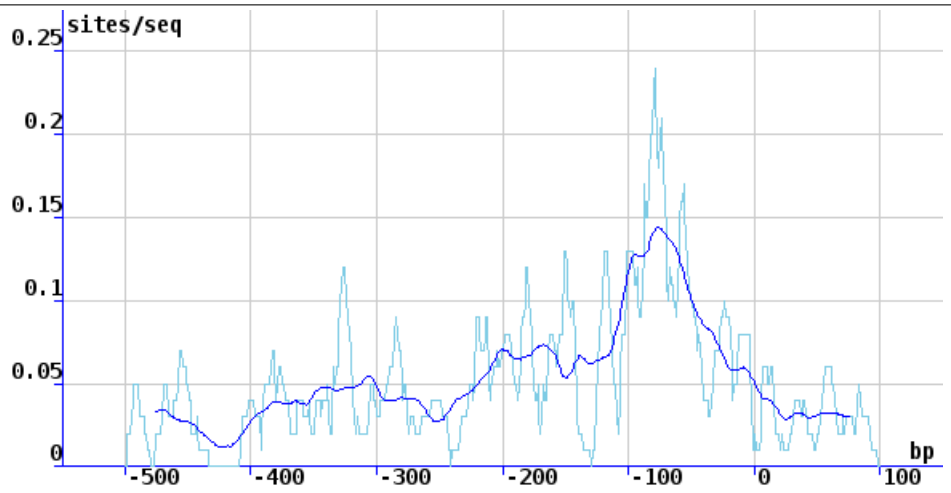
	Matrix name	Yes (sites/1000bp)	No (sites/1000bp)	Yes/No	P-value ▲	Graphs	Matched promoters p-value
☐ (*)	V$NFKB_Q6	14.3067	1.5480	9.2421	7.3475e-07		3.2977e-09
☐ (*)	V$AP2ALPHA_03	29.7139	9.8039	3.0308	1.4078e-06		3.7730e-08
☐ (*)	V$NFKAPPAB50_01	12.1056	1.5480	7.8202	1.1684e-05		4.9252e-07
☐ (*)	V$NFKAPPAB65_01	9.1709	0.5160	17.7733	1.6813e-05		6.0628e-10
☐ (*)	V$NFKB_C	7.3368	0.0000	inf	2.1644e-05		3.2979e-10
☐ (*)	V$IK3_01	13.9398	3.0960	4.5026	6.2795e-05		5.1653e-09
☐ (*)	V$CREL_01	8.0704	0.5160	15.6405	7.4987e-05		6.0628e-10
☐ (*)	V$P50RELAP65_Q5_01	7.3368	0.5160	14.2186	2.0152e-04		1.2966e-07
☐ (*)	V$RELBP52_01	5.1357	0.0000	inf	5.4293e-04		3.4410e-06
☐ (*)	V$CP2_02	7.7036	1.0320	7.4648	6.2760e-04		2.2141e-08

5.2 Detailed graphical report of matrix distribution in promoters

You can inspect the detailed distribution of matrix matches on promoters by selecting one or more matrices in the checkbox column of the output table and pressing the "Show site map for selected matrices" menu link. Graphs will be shown in a new window, whose main component is a new table containing the graphical representation of matches for one promoter per row. The View: **yes set (...)** | background set (...) switch at the top of the frame allows you to view promoters of the main or background set. The [<< Back to summary output] link returns to the sites search result page. The report for the top five over-represented PWMs in Section 5.1.3 is taken as an example below.

The "Sites density" distribution graph (figure below) shows the number of sites lying in specific sequence position (relative to the TSS) divided by total number of sequences. A smoothed graph is also displayed which averages sites density in a window of 50 bp.

Figure 5.17 Sites density distribution graph



5.2.1 Matrix legend

Each matrix from the report is represented by a colored arrow (matrices from the same TF family have identical colors). The length of the arrow corresponds to the length of the site recognized by the matrix. The legend also displays information on site search analysis. By default the legend is placed above the table. After clicking the legend will be placed in the top left corner of the frame when you scroll the page down. By clicking the button you can place it back above the table.

Figure 5.18 Legend of selected PWMs

Legend

		Matrix name	Yes (sites/1000bp)	No (sites/1000bp)	Yes/No	P-value	Graphs	Matched promoters p-value
☑		V$BACH1_01	0.1667	0.0478	3.4857	0.0039		0.0022
☑		V$BACH2_01	0.1500	0.0444	3.3785	0.0070		0.0041
☑		V$IRF_Q6	0.3167	0.1059	2.9910	3.2705e-04		0.0042
☑		V$NFKAPPAB65_01	1.6833	0.5464	3.0805	7.2224e-17		3.2912e-08
☑		V$NFKB_Q6_01	0.9167	0.2391	3.8343	1.7253e-12		9.5913e-09
☑		V$RELBP52_01	1.1333	0.3279	3.4567	1.5155e-13		1.2663e-08

5.2.2 Promoter table

Each line corresponds to the sequence analysed with the site search algorithm, The "Picture" column shows all hits of selected matrices over the whole length of promoter regions (600 bp in our example). Matches are represented by arrows which are located above the promoter line. If the matrix is identified on the positive strand of the promoter sequence, the arrow is pointing to the right. Otherwise, the arrow is pointing to the left. Transcription start sites are indicated by bent black arrows at position 1 of each promoter. The "Number of sites" column shows how many sites of the reported matrices were found during the analysis. Columns with system information (chromosomal position, short description, species name) as well as numerical columns from the main or background gene set depending on the current view are available in the hidden columns list.

Figure 5.19 Sites visualization

	TRANSPro ID	Picture	Number of sites	Gene symbol
☐	HSA_3886_1	-500 -400 -200 1	14	NFKBIA
☐	HSA_14475	-500 -400 -200 1	11	IFIH1
☐	HSA_10123	-500 -400 -200 1	10	EBI3
☐	HSA_12146_1	-500 -400 -200 1	9	NFKB1

Additional annotation columns with system information (chromosomal position, short description, gene accession number) as well as numerical columns with the number of sites for each matrix are available to the right of the graphical display.

Figure 5.20 Sites visualization for selected PWMs: annotation columns

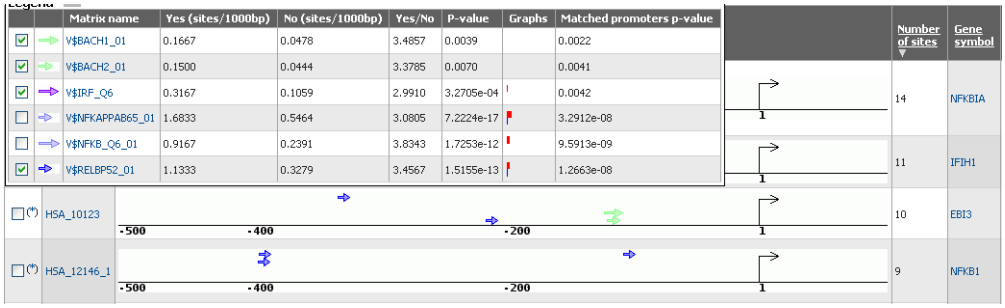
Gene symbol	Description	Probe set	TSS	V$BACH1_01	V$BACH2_01	V$IRF_Q6	V$NFKAPPAB65_01
NFKBIA	nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha	201502_s_at	Chr.14 34943663 (-)	0	0	0	6
IFIH1	interferon induced with helicase C domain 1	219209_at	Chr.2 162883396 (-)	0	0	3	3
EBI3	Epstein-Barr virus induced 3	219424_at	Chr.19 4180520 (+)	1	1	0	4
NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1	209239_at	Chr.4 103641606 (+)	0	0	0	4

5.2.3 Hiding matrices from view

Matrices can be excluded from the representation or included again using the checkboxes in the matrices box. You can see an example where three matrices out of five are excluded. The corresponding matches

in the promoter table are hidden or revealed according to this selection.

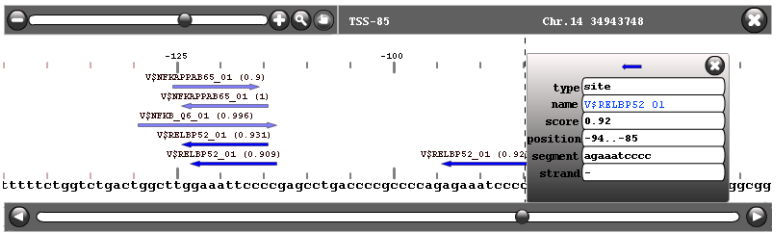
Figure 5.21 Hiding matrices from view



5.2.4 Detailed promoter view

Clicking on a picture of a promoter in the "Picture" area provides a detailed view of the promoter using a site viewer. You can adjust the scale moving the slider in the top left toolbar of the viewer. At the highest resolution, individual nucleotides can be distinguished. Moving the slide on the bottom toolbar, you can move the window of the viewer along the promoter. Detailed information is available by clicking on a specific site. Exact chromosomal coordinates and the TSS of the position denoted with dashed vertical lines, are shown on the top panel. Press the button on the top right corner to close the site viewer.

Figure 5.22 Detailed sequence view via site viewer



Furthermore, the table enables users to select individual promoters in the checkbox column. The "Site search result" menu provides the possibility to save selected promoters as a new gene set clicking on the "Get gene set" link. The "Save report" link saves the current match presentation of one or more matrices in the project tree. A new node for the report appears under the current site search node. The "Get matrices" menu option creates gene sets of matrices checked in the matrix legend.

By clicking on the "Sequence report" link you obtain a detailed text report on the selected promoters. An example of such detailed view for the top promoter is shown on the figure below. The link [<< Back] returns you to the promoters view.

Figure 5.23 Detailed text sequence view

By clicking the "Sites table" link, you can obtain a formatted text report on the selected promoters

in MATCH format. The figure below shows an example of detailed report. Matrix name, entry position, cut-offs, matched sequence and assigned factors are displayed for every matrix hit on selected promoters. The link [<< Back] returns you to the promoters view.

Figure 5.24 Detailed text report

Sites lying on promoter HSA_3886_1
Gene symbol: NFKBIA
Description: **nuclear factor of kappa light polypeptide gene enhancer in B-cells inhibitor, alpha**
TSS: **Chr.14 34943663 (-)**

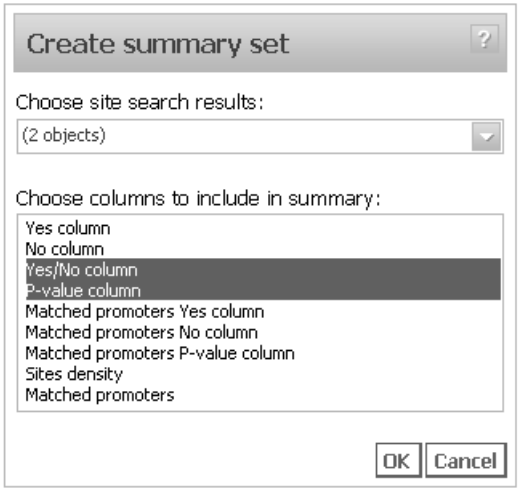
matrix identifier	position (strand)	core match	matrix match	sequence (always the (+)-strand is shown	factor name
V$NFKAPPAB65_01	39(-)	0.867	0.890	GGCAAacccc	RelA-p65
V$NFKAPPAB65_01	121(-)	0.894	0.906	GGGAAacccc	RelA-p65
V$NFKAPPAB65_01	122(-)	1.000	0.960	GGAAAcccca	RelA-p65
V$NFKAPPAB65_01	215(+)	1.000	0.968	aggacTTTCC	RelA-p65
V$NFKAPPAB65_01	376(+)	0.862	0.900	tggaaATTCC	RelA-p65
V$NFKAPPAB65_01	377(-)	1.000	1.000	GGAAAttccc	RelA-p65
V$NFKB_Q6_01	116(+)	0.958	0.927	atcgtGGGAAacccca	IkappaB-gamma,Ikap
V$NFKB_Q6_01	214(-)	1.000	0.928	aaggacTTTCCagcca	IkappaB-gamma,Ikap
V$NFKB_Q6_01	372(+)	1.000	0.996	ggcttGGAAAttcccc	IkappaB-gamma,Ikap
V$RELBP52_01	39(-)	0.986	0.880	ggcaaACCCC	NF-kappaB2-p52:Rel
V$RELBP52_01	121(-)	0.986	0.885	gggaaACCCC	NF-kappaB2-p52:Rel
V$RELBP52_01	377(-)	0.870	0.931	ggaaaTTCCC	NF-kappaB2-p52:Rel
V$RELBP52_01	378(-)	1.000	0.909	gaaatTCCCC	NF-kappaB2-p52:Rel
V$RELBP52_01	407(-)	1.000	0.920	agaaaTCCCC	NF-kappaB2-p52:Rel

Sites lying on promoter HSA_14475
Gene symbol: IFIH1
Description: **interferon induced with helicase C domain 1**

5.3 Summary set of several site search results

ExPlain provides a mechanism to compare several site search results. Using the "Create summary set of several Match results" option from the "Analyze" menu, launch the dialog displayed on Figure 5.25. Select several results of site search analysis described in Chapter 5, «Transcription factor site search analysis». Select columns that should be present in the output. Note that the "Sites density" column will appear in the output dataset if MATCH results without background set were present among the selected results.

Figure 5.25 MATCH summary dialog



When you press the OK button, a process of summary set creation will be started. The output set will contain matrices from all marked site searches and the selected columns grouped by MATCH results. The figure below shows a summary of results from Figure 5.3 and Figure 5.16.

Rows with the most different and most similar p-values are marked in the "sim/diff" column.

Figure 5.26 MATCH summary set

Result »			vertebrate_all 600 SUP (Up HUVEC GSE2639 example, NC HUVEC GSE2639 example)		vertebrate_all 600 SUP filtered by known_sites (Up HUVEC GSE2639 example, NC HUVEC GSE2639 example)		
	Gene symbol	Matrix name	Yes/No	P-value	Yes/No	P-value	sim/diff
	CEBPD	V$CEBPDELTA_Q6	1.1528	0.0282	2.7015	0.0286	sim
	CHURC1	V$CHCH_01	1.0834	0.0117	1.5136	0.0335	sim
	HMGA1, HMGA2	V$HMGIY_Q3	1.1753	0.0224	3.3177	0.0354	sim
	IKZF1	V$IK3_01	1.2453	2.3650e-05	4.5026	6.2795e-05	sim
	IKZF1	V$IK1_01	1.1916	0.0051	3.4125	0.0050	sim
	NFIC	V$NF1_Q6_01	0.8167	0.0200	0.1185	0.0232	sim
	NFIC	V$NF1_Q6	0.8487	0.0456	0.2031	0.0313	sim
	NFKB1, NFKB2, REL, RELA	V$NFKAPPAB_01	2.0211	3.2929e-12	7.1093	0.0010	diff
	NFKB1, NFKB2, REL, RELA	V$NFKB_C	2.1047	2.7717e-10	inf	2.1644e-05	diff
	NFKB1, NFKB2, REL, RELA	V$NFKAPPAB65_01	3.0805	7.2224e-17	17.7733	1.6813e-05	diff

5.4 Sites search theoretical background

5.4.1 The TRANSPro database

Numerous analyses require actual promoter sequences, including promoter prediction tools and analyses of gene regulation in gene expression experiments. Being aware of the existence of several non-interchangeable definitions of the term promoter, we use it here for those sequences located in the vicinity of a transcription start site (TSS).

The extraction of reliable promoter sequences is usually a difficult task demanding a huge amount of tedious handwork. To relieve this situation, the TRANSPro database was introduced as module to the TRANSFAC suite adding extensive annotation for upstream (5′) sequences of human, mouse, and rat genes. The emphasis is made on the elements involved in gene regulation.

Underlying sequence databases. TRANSPro is based on Genomic Sequence Assemblies from the international sequencing consortia in the Ensembl database. Promoter sequences are extracted only for those genes for which both an Entrez Gene ID and a nomenclature accession number (HGNC for human, MGI for mouse, and RGNC for rat) are defined.

5.4.2 Computational definition of transcription start sites

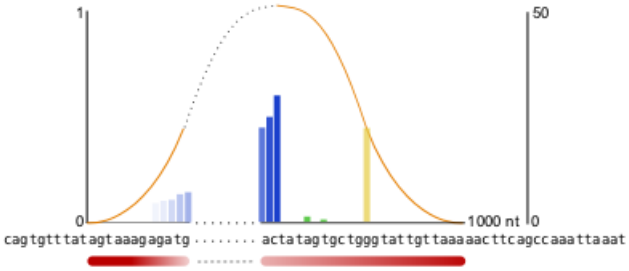
TRANSPro integrates TSS data from EPD [24], DBTSS [25] and Ensembl to derive virtual TSSs as reference points. DBTSS entries are assumed to be the first nucleotide of the one-pass mRNA sequences, whereas Ensembl TSSs resemble the first nucleotide of the 5′-most exon of an Ensembl mRNA model. Consequently, the collection of TSSs for some genes may be scattered over a genomic sequence segment spanning several thousand nucleotides, sometimes even more than 100 kb.

In order to acquire a reasonable number of TSSs for a given gene, an algorithm was designed to cluster TSS locations from the source databases (evidence points) around virtual TSSs which are then used as reference. A window of 1000 nt length is slid along the entire sequence fragment containing all TSSs and a clustering score is calculated by summing up contributions from each evidence point within the window weighted according to their source. A score of 1 is added for each DBTSS or Ensembl entry, whereas EPD entries receive a score of 50 assuming higher reliability due to manual expert annotation. Scores of individual evidence points are multiplied by a distance factor ranging from 1 at the center of the window to 0 at the two outer positions. In between, the distance factor is computed by a cosine function with lower values the greater the distance to the window center. Peaks of the score histogram are regarded as virtual TSSs if the corresponding sequence window contains at least 5% of all evidence points, so that multiple virtual TSSs per gene are well accepted. However, for some genes only a handful of evidence points are available, yielding several virtual TSSs, yet, with few references. To accommodate such cases, annotation restrains to the 5′-most virtual TSS for genes with less than 20 evidence points.

Virtual TSSs defined by the method described above form the basis for the extraction of TRANSPro promoter sequences in a fully automatic fashion. Whenever conflicts or inconsistencies occur the respective gene is excluded from the TRANSPro database. The components of the virtual TSS definition method of TRANSPro are summarized in Figure 5.27. The window of 1000nt length signified by the red bars is slid along the genomic sequence segment and the sum of scores of all evidence points within the window is computed. Ensembl and DBTSS entries receive a maximal score of 1, whereas EPD TSSs obtain a maximal score of 50. Scores are multiplied by a penalizing distance factor, where a score in the

center of the window is multiplied with 1 and the factor diminishes to 0 according to a cosine function the greater the distance to the center. In the figure, 3 evidence points within the window are indicated by green and yellow bars. The cosine is exemplified by the orange curve. Finally, a sum-of-evidence-scores histogram peaking at the current window position is shown by blue bars.

Figure 5.27 TSS definition in TRANSPro



5.4.3 Cut-off values

In the MATCH, cut-off values are separately defined for core and matrix similarity values [5]. The matrix similarity is a score that describes the quality of a match between a matrix and an arbitrary part of the input sequences. Analogously, the core similarity denotes the quality of a match between the core sequence of a matrix (i.e. the five most conserved positions within a matrix) and a part of the input sequence. A match has to contain the "core sequence" of a matrix, i.e. the core sequence has to match with a score higher than or equal to the core similarity cut-off. In addition, only the matches that score higher than or equal to the matrix similarity threshold appear in the output. For the minFP, minFN, and minSUM cut-offs, first the core similarity score is calculated, and then the matrix similarity score is calculated for the selected positions according to the following equation.

$$mat\_sim = \frac{\sum_{i=1}^{L} I(i) f_{i,b_i} - \sum_{i=1}^{L} I(i) f_i^{\min}}{\sum_{i=1}^{L} I(i) f_i^{\max} - \sum_{i=1}^{L} I(i) f_i^{\min}}$$

where f_{i,b_i} is the frequency of nucleotide b at position i of the matrix with width L, $f_i^{\min}$ the frequency of the rarest occurring nucleotide in position i, and $f_i^{\max}$ the frequency of the most frequent occurring nucleotide in position i. The information vector I(i) describes the conservation of nucleotide B in position i of the matrix:

$$I(i) = \sum_{B \in \{A,C,G,T\}} f_{i,B} \ln(4 f_{i,B})$$

, i = 1, 2, ..., L

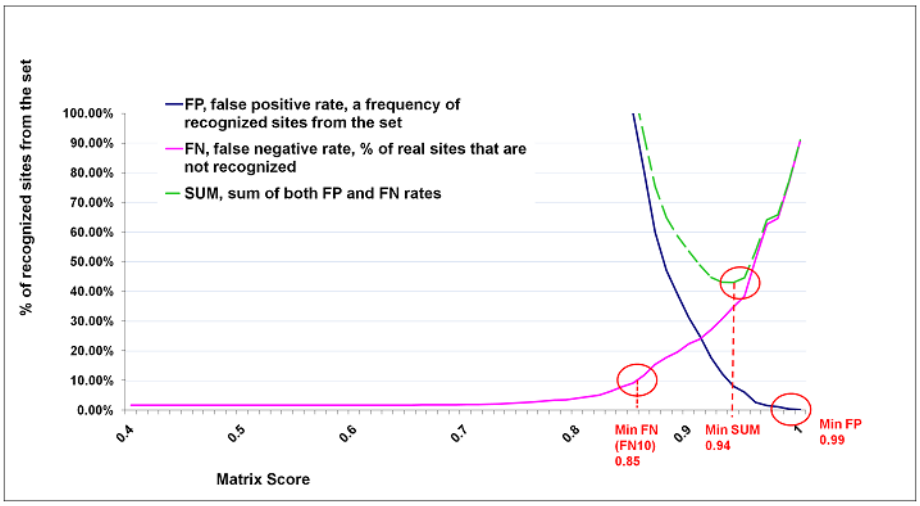
Cut-off to minimize false negative matches (minFN): The false negative rate MatchTM obtains with a matrix was measured on known genomic binding sites for the transcription factors associated with that matrix, as far as such sites are available. In case a sufficient number of genomic binding sites (less than 10) were not available, SELEX sites or sets of generated oligonucleotides were used for estimating the cut-offs to minimize the false negative rate, using actual weight matrices to calculate the probability of a nucleotide occurring at a certain position of a binding site. For each matrix we applied the MatchTM algorithm to these test sequence sets without using any matrix similarity cut-offs. Then we set the cut-off to a value that provides recognition of at least 90% of oligonucleotides. We decided to tolerate an error rate of ten percent. We call this set of cut-offs minFN cut-offs. Applying the minFN cut-offs, the user will find most genomic binding sites, but in this case a high rate of false positives should be taken into account as well. The minFN cut-offs are useful for the detailed analysis of relatively short DNA fragments.

Cut-off to minimize false positive matches (minFP) : In order to estimate this cut-off, which will reduce the number of random sites found by MatchTM, we applied the MatchTM algorithm to promoter

sequences from TRANSProTM 2.1. (in the Matrix Generation tool, exon sequences are used). The score that gives 1% of hits in these sequences relative to the number of hits received when using the minFN score (calculated above) is defined as minFP. When a minFP cut-off is applied for searching a DNA sequence, the algorithm will find a relatively low number of matches per nucleotide. In the output the user will only find putative sites with a good similarity to the weight matrix; however, some known genomic binding sites could not be recognized. This kind of cut-off is useful, for example, for searching the most promising potential binding sites in the extended genomic DNA sequences.

Cut-off to minimize the sum of both error rates (min SUM): We compute a sum of both error rates to find cut-offs that give an optimal number of false positives and false negatives (Figure 5.28). To do so, we compute the number of matches found in promoter sequences for each matrix using a cut-off allowing 10% of false negative matches (minFN10). This number is defined as 100% of false positives. The sum of corresponding percentages for false positives and false negatives is then computed for every cut-off ranging from minFN10 to minFP. We refer to the cut-off that gives the minimum sum as minSum cut-off.

Figure 5.28



5.4.4 Calculation of theoretical P-values for TFBS predictions

In the following, we describe an extension of the standard method of exact calculation of P-values for PWM scores (MSS, Section 5.4.3) concerning calculation of score densities considering a search over both sequence orientations and according to a 1st order Markov model. Let $Q = (q_{ik})$ be the W x 4 score matrix with i = 1 .. W vectors, also called site positions, with k = 1 .. 4 scores for the residues $R = (r_k) = (A, C, G, T)$. This matrix is used to score an alignment of the TFBS profile with a sequence segment of length W. The score function is typically additive, so that its value is calculated by summing the site position scores of the sequence residues. Further let $\pi = (\pi_k)$ be a mononucleotide background model of residue frequencies. Usually, one considers both sequence orientations or, equivalently, both orientations of the PWM. A method for P-value calculation should take this into account and determine the probability that either score (forward or reverse orientation) is above a threshold. The reverse PWM is defined in the following equations.

$$Q' = (q'_{ik}) \quad \text{with } i = 1 \ldots W;\ k = 1..4$$

$$q'_{ik} = q_{(W+1-i)(5-k)}$$

Given forward and reverse orientation, there is a pair of scores (q,q') for each position of an alignment and subsequently a pair of scores (s,s') for the sequence segment. As shown in [11], the probability distribution of PWM scores can be efficiently calculated by convolution. The latter can be extended to calculate the joint probability density f(s,s'). The equation below gives the joint probability density of scores up to position i of the alignment by convolution of g(q,q') and h(s-q,s'-q'), where g is the function of score pairs at position i of Q and Q' and h is the joint probability density of score pairs up to position i-1.

$$f(s_i, s'_i) = [g * h](s_i, s'_i) =$$

$$= \sum_k g(q_{ik}, q'_{ik}) * h(s_i - q_{ik}, s'_i - q'_{ik}) \quad \text{with } k = 1..4$$

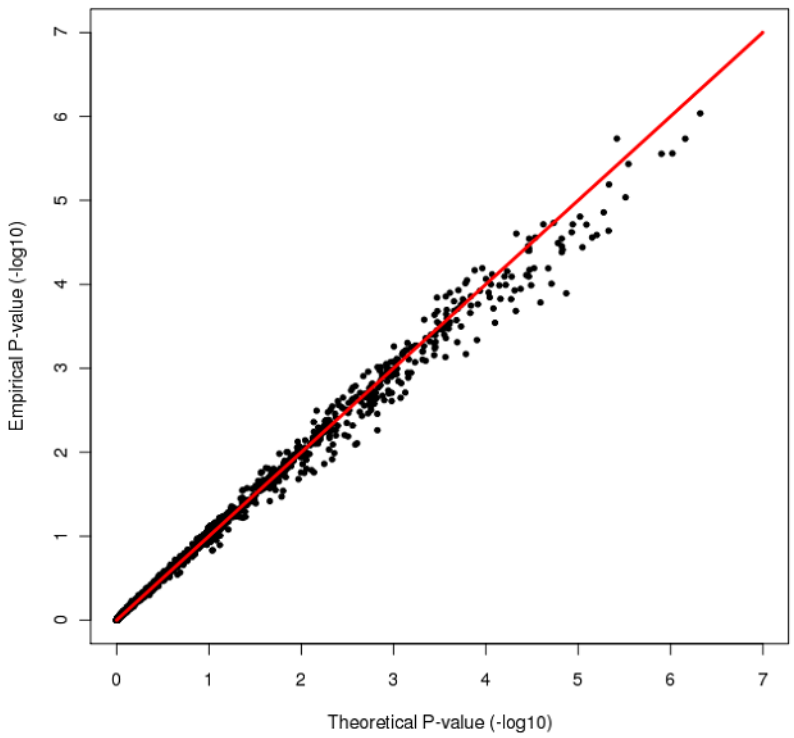
The cumulative probability $P(S \geq t) \cup P(S' \geq t)$ that determines selection of score threshold t is derived from the joint probability density of scores. We further extend the standard method by applying a dinucleotide background model. By conditioning score functions on the terminal residue j associated with a score pair (si,s'i), the convolution can also accommodate higher order background models.

$$f(s_i, s'_i) = [g_j * h_k](s_i, s'_i) =$$

$$= \sum_k g(q_{ij}, q'_{ij}) * h_k(s_i - q_{ij}, s'_i - q'_{ij}) \quad \text{with } j, k = 1..4$$

The accuracy of theoretical P-values using a uniform background model is demonstrated in Figure 5.29. Sites were predicted at all score thresholds in 1000 real promoter sequences and afterwards the empirical prediction rate corresponding to each score threshold was compared to the theoretical value. In case of perfect correspondence between theoretical and empirical P-values, points fall on the diagonal (red line on the figure).

Figure 5.29



5.4.5 Sites search optimization with F-Match algorithm

The F-Match algorithm compares the number of sites found in a query sequence set against the background set. It is assumed, if a certain TF (or factor family), alone or as a part of a cis-regulatory module, plays a significant role in the regulation of the considered set of promoters, then the frequency of the corresponding sites found in these sequences should be significantly higher than expected by random chance. Often, the stringency of the interaction of this TF with their target sequences in the considered promoters is not known, leading to the uncertainty in setting thresholds on the site searches using the MATCH program.

F-Match carefully evaluates the set of promoters, and for each matrix tries to find two thresholds: one, th-max, which provides the maximum ratio between the frequency of matches in the promoters in

focus (query set) and background promoters (background set) (over-represented sites); and the second threshold, *th-min*, that minimizes the same ratio (underrepresented sites). As a result, for each weight matrix we obtain a set of predicted K sites and M sites in the both promoter sets with the corresponding matrix scores.

The F-Match algorithm makes an exhaustive search through the space of all scores observed in the sequence sets. Each observed score is taken as a threshold *th* and the program computes the number of sites *k* found in the main promoter set and number of sites *m* found in the background promoter set. Then, the expected number of sites in the main set to be observed in the case of even distribution of sites between two sets will be:

quation 5.4.1

$$k_{\exp} = n \cdot f, \quad n = k + m, f = \frac{|main\ set|}{|main\ set| + |background\ set|}$$

and assuming a binomial distribution of the sites between two sets, we can calculate the *p-value* of finding the observed number of sites and higher, for over-represented matches, or lower, in the case of under-represented matches

quation 5.4.2

$$if\ k \geq k_{\exp},\ p - value(+) = \sum_{i=k}^{n} C_i^n \cdot f^i \cdot (1-f)^{n-i},$$

$$if\ k < k_{\exp},\ p - value(-) = \sum_{i=1}^{k} C_i^n \cdot f^i \cdot (1-f)^{n-i}$$

giving the p-value of over- and under-representation of matches in the main promoter set.

For a given significance level *p* (e.g. *p = 0.001*), F-Match finds such thresholds *th-max* and *th-min* that maximizes and minimizes, respectively, the ratio k/k_{exp} provided that the *p-value* $< p$. If the required significance level cannot be reached for a given matrix, this matrix will not be considered.

5.4.6 The P-Match algorithm

P-Match combines pattern matching and weight matrix approaches, thus providing higher accuracy of recognition than each of the methods alone. The algorithm is based on simultaneous use of a positional weight matrix (PWM) and a set of aligned TF binding sites used to construct this matrix.

The P-Match search algorithm computes *d*-score value which measures similarity between a subsequence *X* of the length *L* in DNA and a given TF site *S* from the site set. The *d*-score is calculated using weights of the nucleotides in the individual positions of the site taken from the corresponding weight matrix:

quation 5.4.3

$$d = \frac{\Delta w - \sum_{i=1}^{L} w(i, B(X_i)) - w(i, B(S_i))}{\Delta w}$$

$$\Delta w = \sum_{i=1}^{L} \max(B(ACGT) w(i, B)) - \sum_{i=1}^{L} \min(B(ACGT) w(i, B))$$

where *B(Xi)* and *B(Si)* are the nucleotides in *i*th position of the subsequence *X* and the site *S*, respectively. The *d*-score ranges from 0.0 to 1.0, where 1.0 denotes an exact match of the nucleotide weights

of the site S to the corresponding weights of the sub-sequence X. Similar to the Match algorithm, we compute two separate *d*-scores: *d-matrix* for the whole site and *d-core* for the core positions of the site, which are the five most conserved positions in the alignment. In the search, we choose two independent cut-off values for these two *d*-scores (in each site set, the cut-offs are the same for all sites in the set). Only those matches are reported whose both *d*-scores dmatrix and dcore exceed the corresponding cut-offs.

The sets of aligned sites contain many sites that are similar to each other, out of several overlapping matches, those with the highest *d*-score will be outputed. In addition, the high similarity of a match in the sequence under study to an existing known site in a promoter of another gene can give an idea about the function of the found potential binding site.

Three different *d*-score cut-offs are precalculated for each TRANSFAC matrix that has the site alignment set attached to it:

(i) to minimize false negative rate (under-prediction error);

(ii) to minimize false positive rate (over-prediction error);

(iii) to minimize the sum of both errors.

5.4.7 The Patch algorithm

The Patch algorithm is designed for searching potential binding sites for transcription factors (TF binding sites) in any sequence which may be of interest. The patterns, which Patch uses for searching, are TF binding sites of the TRANSFAC Professional database and the consensus sequences of weight matrices of TRANSFAC Professional.

Number of parameters used to limit the results of the search.

Minimum length: This parameter specifies the minimum length for sites which are shown in the Patch outputs. Using the default value 10, only sites longer than or equal to 10 will appear in the output. Please note that the maximum number of mismatches allowed also influences the minimum sequence length. For additional information please refer to the next paragraph.

Maximum number of mismatches: It would be more precise to call this parameter the maximum number of local mismatches, as it specifies how many positions may differ when comparing a binding site (search pattern) with some part of the input sequence. A match between the whole site and the input sequence has been found, if the actual number of (local) mismatches is lower than or equal to the maximum number of (local) mismatches. Please be careful when selecting this value. The maximum length of the sites searched for (search patterns) is restricted to 2*maximal_number_of_mismatches+1. That means, if you set the maximal number of mismatches to 5, PatchTM searches only for sites which are longer than 11bp. All shorter ones will be ignored. The default value for this parameter is 0.

Mismatch penalty: When comparing a binding site (search pattern) with some part of the input sequence, each mismatching position will receive a mismatch penalty. This penalty value will have a negative influence on the overall score for the match between the whole site (search pattern) and the input sequence. Each matching nucleotide receives a bonus weight of 100. So, the default value for the mismatch penalty is also 100, and the negative influence of a mismatching position corresponds to the positive influence of a matching position. If you reduce the mismatch penalty, you will receive high scoring sites containing mismatches in the Patch output. If you increase this parameter, high scoring sites are not likely to contain mismatches.

Lower score boundary: The lower score boundary is a cut-off, defining which matches between a site (search pattern) and the input sequence will be listed in the output. The score which is estimated for every match has to be higher than or equal to this cut-off. The default value for the lower score boundary is 87.5.

5.4.8 The Seeder algorithm

Seeder is a discriminative seeding DNA motif discovery algorithm designed for fast and reliable prediction of cis-regulatory elements in eukaryotic promoters.

The motif search starts by enumerating all nucleotide combinations (words) of a given length, usually six. For each word, it calculates the Hamming distance (HD) between the word and its best matching

sub-sequence (the substring minimal distance — SMD) in each sequence of a background set. This data is used to produce a word-specific background probability distribution for the SMD. For each word, it then calculates the sum of SMDs to sequences in a positive set. The P-value for this sum is calculated using the word-specific background probability distribution. The word for which the P-value is minimal is retained, and a seed PWM is built from the closest matches to this word found in every positive sequence. The seed PWM is extended to full motif width and sites maximizing the score to the extended PWM are selected, one in each positive sequence. A new PWM is built from those sites and the process is iterated until convergence, or a maximum number of iterations is reached.

Input data and parameters: the algorithm takes as input a set B = {B1 , ..., Bm } of m background sequences of length L, a set P = {P1 , ..., Pn } of n positive sequences of length L, the length k of the motif seed and the length l of the full motif to discover. The SMD distance d(a,b) between a short nucleotide sequence a and a longer sequence b is the minimal HD between a and a |a|-length substring of b .

Background model: a discrete random variable Y (a) is associated with each word a of seed length k, corresponding to the SMD between a and a randomly selected background sequence from B. This a-specific distribution function is obtained empirically from B; for each word a, one sets ga (y) = Pr[Y (a) = y] = |{Bi : d(a,Bi) = y}|/m, for y = 0, ..., k.

Seed position weight matrix: for each word a, the sum of SMDs to the positive sequences S(a) = $\sum_j$ d(a, Pj) is computed. Under the background model, the distribution function of this sum of n independent and identically distributed random variables is gn*a (y), the n-fold self-convolution of ga (y). The P-value (p) for word a with sum S(a), which is the probability of obtaining a sum lower or equal to S(a) under the assumption that Pj's are random in respect to a, is:

quation 5.4.4

$$p(S(a))= \sum_{y=0}^{S(a)} g^{n*}{}_a (y)$$

The word a* for which the P-value p(S(a)) is minimal is retained. For each positive sequence in P, the set of one or more subsequences of length k having the SMD to a* are retained. A PWM P0 is built from this set of selected subsequences using standard procedures and pseudocounts proportional to $\sqrt{n}$, with the modification that when a sequence contains more than one match, each match (sub-sequence) weight is reduced proportionally. The sub-sequence associated with the highest score to P0 is retained in each sequence, and the seed PWM Ps is built from this optimal set of n sub-sequences. Full length motifs: the original seed PWM Ps is of width k, probably smaller than the full motif width. It is extended to full motif width l by adding null weights at (l - k)/2 positions upstream and downstream. The full length PWM is then refined by iterating the following process. (i) Sites (one per sequence in P) maximizing the score to the extended weight matrix are selected and (ii) a revised full length PWM is built from those sites. This process is repeated until convergence (i.e. the sites maximizing the PWM score are fixed in all sequences) or for at most a default number of 10 iterations, which are often sufficient for the convergence of significant seeded motifs. For each motif predicted, a list of 4k P-values is generated thus prompting for a multiple testing correction. This is carried out by generating a list of q-values from the list of P-values associated with words of seed length k, using a general algorithm for estimating q-values. The statistical significance of a motif is evaluated with the q-value of the sum S(a*), which is the expected proportion of false positives incurred when calling the sum significant (i.e. not likely to have occurred if the positive sequences were randomly selected).

Reference: F.Fauteux, M. Blanchette, and M.V. Stromvik, Bioinformatics 2008 (24), 2303-2307.

Chapter 6

Composite module analysis and models

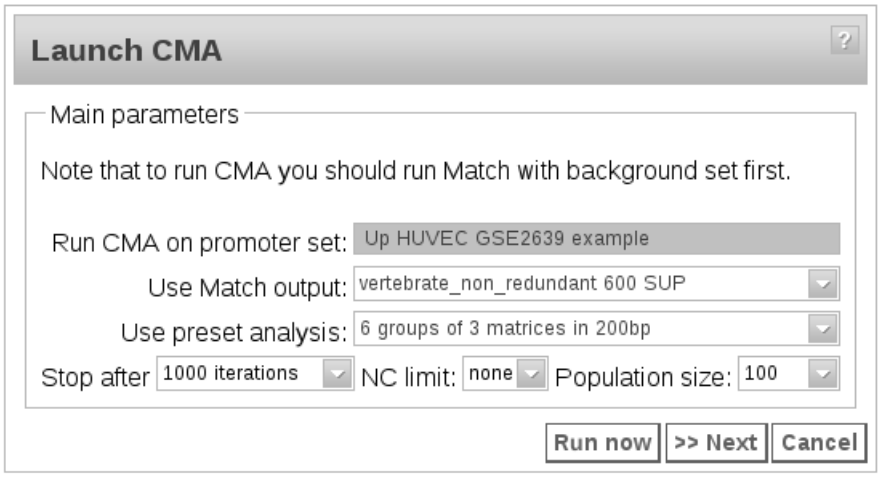
The composite module analysis allows you to derive promoter models for your set of target promoters. A promoter model characterizes promoters by a set of binding motifs and by rules for their arrangement as single motifs and pairs. Promoter models are optimized with a Genetic Algorithm for maximal discrimination between target promoters and a control set. Models obtained through this method can be used to classify other promoter sequences or to investigate the signaling network upstream of the transcription factors binding to the motifs present in the promoters of a gene set.

6.1 The CMA interface in ExPlain

The "CM: Composite Models (genetic algorithm)" link from the "Analysis" menu launches a dialog window providing an interface to the CMA program. The second way to activate the CMA dialog is to press the button on the toolbar. In the following example we show how to use the ExPlain interface to obtain CMA models for your target promoters. The Section 6.6 section explains how to use CMA models for promoter classification.

6.1.1 The CMA dialog window

Figure 6.1 CMA dialog window



The CMA dialog window allows for adjustment of general parameters. CMA retrieves information about binding site predictions and corresponding PWMs from a set of promoters, thus if you choose a precomputed MATCH output in the "Use Match output" field, the respective promoter set will be chosen automatically. Note, that you can only select the output of an analysis run against a background set. The "Use preset analysis" field provides the list of preliminary saved model constraints and GA parameters. Later, in this chapter we will discuss how to save your own preset parameters.

You can further adjust the number of iterations or running time, the NC limit, and the population size for the GA. The NC (No Change) limit specifies the number of iterations after which the program is stopped if the best fitness has not improved by more than 0.0001 during that interval. For instance, if CMA was set to use 20 minutes with a NC limit of 200, the program can stop before reaching the time limit if no improvement was achieved over the last 200 iterations. Finally, the population size is the number of individual model solutions the algorithm can take into account during each iteration.

ASTUCE

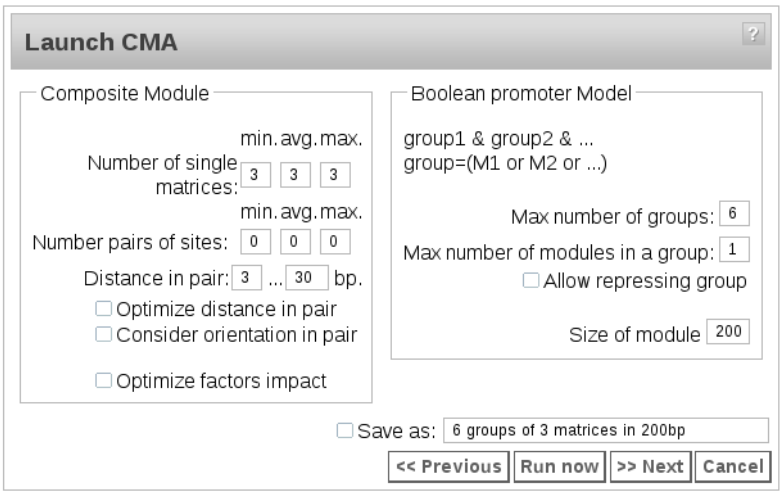
The population size and the running length parameters have great effect on the model quality you can obtain. You should set both to high values if, for instance, you have a large number of target promoters or consider complex assemblies of single PWMs, PWM pairs and groups. In general, when the complexity of the optimization task increases these two parameters can be used to optimize the algorithm toward finding the best model.

Press Run now button to start the analysis after you set up all parameters.

6.1.2 The CMA advanced dialog window

The advanced options of CMA are available through the >> Next button.

Figure 6.2 CMA advanced dialog window


Launch CMA
Composite Module
min. avg. max.
Number of single matrices: 3 3 3
min. avg. max.
Number pairs of sites: 0 0 0
Distance in pair: 3 ... 30 bp.
Optimize distance in pair
Consider orientation in pair
Optimize factors impact
Boolean promoter Model
group1 & group2 & ...
group=(M1 or M2 or ...)
Max number of groups: 6
Max number of modules in a group: 1
Allow repressing group
Size of module 200
Save as: 6 groups of 3 matrices in 200bp
<< Previous
Run now
>> Next
Cancel

The Composite Module dialog contains fields to adjust the CM components of promoter models. You can express your assumptions about the abundance of single matrices and matrix pairs by setting minima and maxima as well as the average number of these components in the dialog window. For pairs, you can also adjust the spacer range in the "distance in pair" field. Furthermore, you can require CMA to optimize the allowed spacer range (in steps of 6 bp) and to take specific orientation into account. You can optimize factors impact by checking corresponding option. In this mode weights will be assigned to the factors according to their contribution toward widening the difference between the positive and negative sets.

ASTUCE

The distance optimization and orientation options can help to make your models more discriminative. For small target set sizes, however, the inclusion of these constraints should be carefully considered, if you plan to use your models for classification of new promoters. Small target sets, can promote overfitting, because they are less likely to be representative of the whole promoter family you wish to model.

ASTUCE

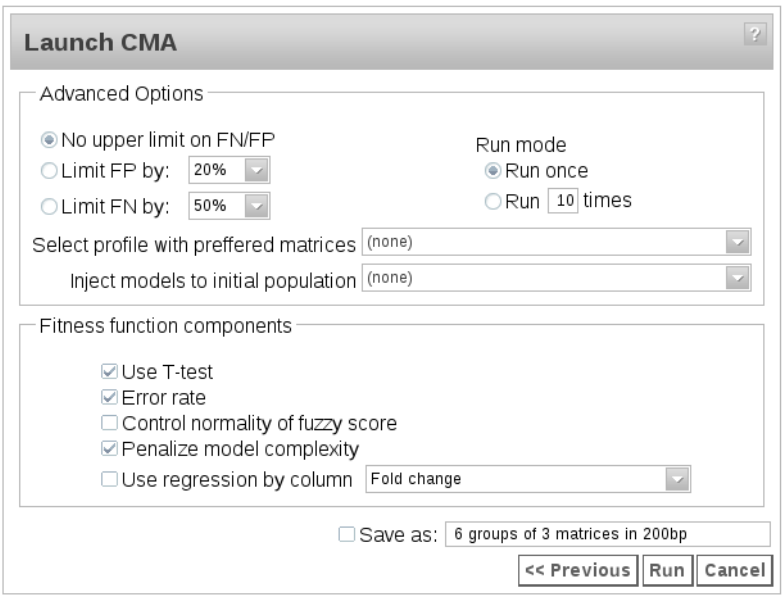
To require a fixed value in single matrix and pair number specifications, you can set minimum, average and maximum for this value, as well as for minimal and maximal spacer lengths in the distance field.

The Boolean promoter model dialog provides for options that apply to more general features of CMA models. You can specify the maximal number of groups and the maximal number of CMs per group as well as the length of the CM sliding window. Furthermore, you can allow for a repressing module. The repressing module is included in the model with the logical operator NOT. Scores of the repressing module should be inversely correlated with expression values used for separation of the two input sets. This option makes sense for some specific analyses, such as comparison of up-regulated and down-regulated promoters.

The advanced dialog allows you to save a defined combination of parameters. You should mark the checkbox and specify the name of a new preset. When you launch CMA by pressing the Run now button, the preset will be saved. It will appear in the project tree inside the "CMA" subfolder of the "Preset" folder.

Use << Previous to return to the previous step. The last set of adjustable options is available on the third CMA dialog screen, appearing after the >> Next button is pressed.

Figure 6.3 CMA advanced dialog window



The FP/FN field allows you to customize the false positive and false negative weights in the E component of the fitness function. You can increase the value for either FP or FN to make the corresponding error rate affect the fitness of models more strongly. If you wish to give more importance to suppressing

false positive errors, you set the FP restriction to a higher level. This will simultaneously lower the FN restriction.

You can further choose to run the program once or any specified number of times. For stochastic optimizers it is often recommended to run the program multiple times, compare results of all runs afterwards and settle on a consensus among different runs. GAs work stochastically and may deliver different results even when invoked with the very same parameters. This is especially true for large, complex optimization tasks, where the algorithm is more likely to end up with different solutions. On the other hand, the runtime is typically increased for larger assignments generally (larger populations, more iterations).

ASTUCE

For large assignments, you can first increase the population size and optimization runtime and afterwards examine not only the best but perhaps the 20 best models, to see whether these deviate strongly from each other, for instance, in the motifs that compose them.

It is possible to determine set of PWMs, which are preferred to be in the modules found. The CMA will search for modules containing at least one of the preferred PWMs as a single matrix, and at least one of the sites for a pair of sites should appear in the selected profile as well.

Finally, you can also customize the five fitness components which are described above. You should select desired components using checkboxes. For instance, if you find that normality of promoter scores is not important at all, you might deselect the corresponding line to make CMA ignore this model property.

To save the combination of parameters, select the save option and edit a name for the parameter set, as it was described above.

6.1.3 The CMA output

We will consider the MATCH output created in Section 5.1.3. After having searched up-regulated genes with non-changed genes as a negative set, we would like to see whether we can find reasonable composite elements, i.e. pairs of factors that may act synergistically on promoters of up-regulated genes. Composite module parameters allow from zero to three single matrices with two matrices on average and one matrix pair. Instead of the CMA default spacer range of three to thirty, we allow for a range of zero to forty and let the program optimize the pair spacer. We further allow for one group each containing maximally five non-mutually exclusive composite modules with a window length of 100 nucleotides. Other parameters take their default values. We let the program optimize for maximally 20 minutes with an NC limit of 200 iterations and a population of 1000 solutions.

When you select a node of a running CMA process in the project tree, you can inspect the current status of iterations, best model and fitness. The ”Processing” field provides a progress bar followed by the portion of iterations already processed as well as the status of the NC limit. Below the ”Processing” field, the best model calculated up to the moment is shown. You can save current model using ”Save model” menu link, or view the extended description clicking on [Verbose display] link above the model. Below the model you are provided with the maximal fitness in the current population (the fitness of the best model(s)) and a plot of the best fitness development up to the current iteration.

The CMA process can be stopped by the button. Then the calculation will be successfully finished and the best model calculated up to that moment will be taken as the result.

Parameters of the search as well as the fitness plot of the process can be recalled by expanding additional information (see Section 1.5.2). You may also modify the stop condition while CMA is running by expanding [+ Parameters] block and changing parameters there.

The ”Stop after” and ”NC limit” parameters have the same meaning as in the CMA dialog. New parameters will take effect only after

button is pressed. Note that if you set the running time to be less than what the CMA has already run, it will stop after the current iteration.

After the calculations are completed, you will see a comprehensive report about the best model found. The presentation contains a graphical description of the model with its performance on the dataset and a table with details about the fitness calculation.

Figure 6.4 CMA progress information

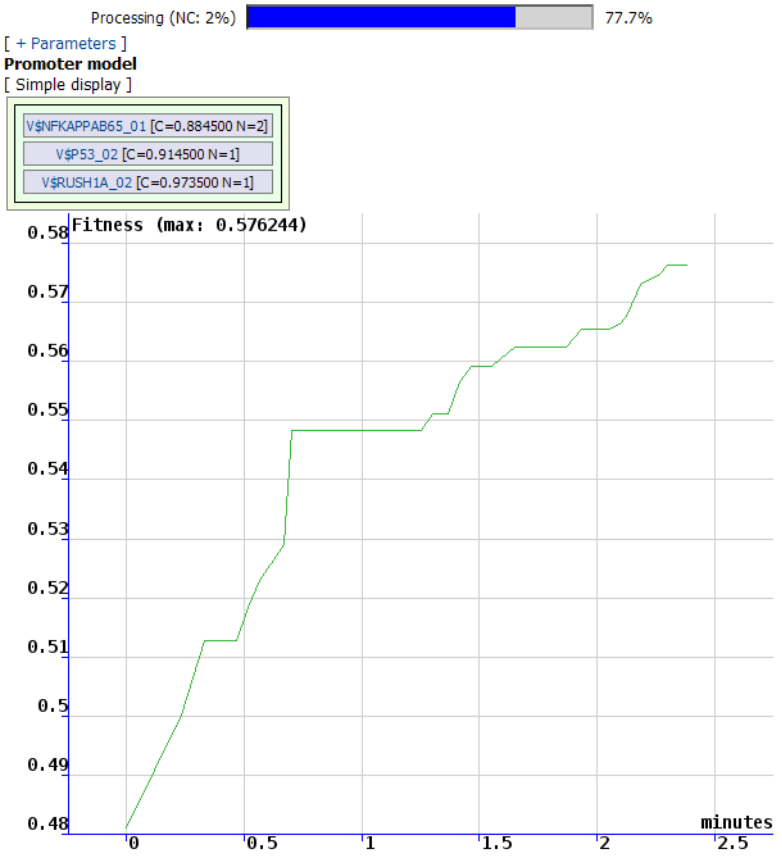
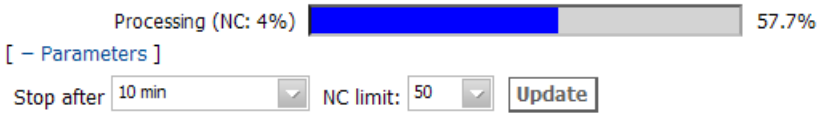


Figure 6.5 Changing CMA stop condition on the fly



The model description is expanded in Figure 6.6. In this mode, the graphic also displays the matrix cut-offs (C=0.972500 for the first matrix) and the number of matrix matches expected in the module (N=1). To switch between expanded and simple display click the link above the model. Underneath the description, false positive (FP: 13.07%) and false negative (FN: 30.95%) frequencies on the dataset as well as the CMA overall cut-off (Overall cut-off: 0.112918) for the model are given.

The "Value" row of the "Goal function calculation" table shows the performance of the model with regard to individual fitness function components. The "Weight" row contains weights of the individual components. These weights resemble the values set in the "Advanced GA options" dialog of the application frame normalized to sum to 1 (in our example the components T, E and P were selected). The total fitness of the model is computed from the sum of the weighted fitness values as demonstrated in the bottom row of the table named "Weighted value".

The "Expression-score", "Yes-No" and "Sites density" distribution plots are available through [+ Expression-score and Yes/No distribution] button.

The "Expression-score" distribution indicates how well scores from the promoter model fit to input sets (main and background set). In our example, the two sets were selected separately and CMA assigned the expression value of 1 to the main set, and -1 to the background. The separation between positive and negative set is indicated by the horizontal line in the plot. In our example, the separation line is congruent with the horizontal axis at the zero expression level.

The "Yes-No" distribution plots observed frequencies of model scores for the main (red) and background (blue) set. This presentation reflects how well the model discriminates between promoter sequences of the two sets.

Figure 6.6 Top section of a promoter model report

Promoter model

[Simple display]

V$CREL_01 [C=0.972500 N=1]

V$IRF_Q6 [C=0.979500 N=1]

<-V$CETS168_Q6 [C=0.841500]-> [19..25] <-V$HAND1E47_01 [C=0.821500]-> [N=3]

P-Value: 2.3226e-18
FP: 13.07%
FN: 30.95%
Overall cutoff: 0.112918

Goal function calculation

Elements	R	T	E	N	P	Total
Value	1.000000	0.651198	0.779879	0.003198	0.511099	
Weight	0.000000	0.333333	0.333333	0.000000	0.333333	1.000000
Weighted value	0.000000	0.217066	0.259960	0.000000	0.170366	0.647392

Figure 6.7 "Expression-score" distribution

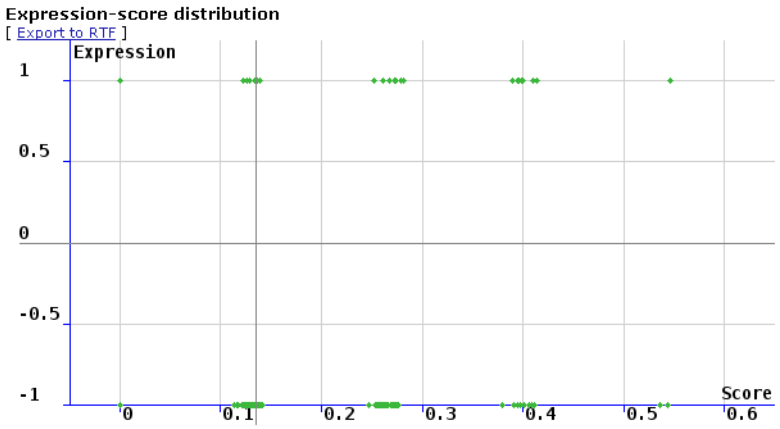
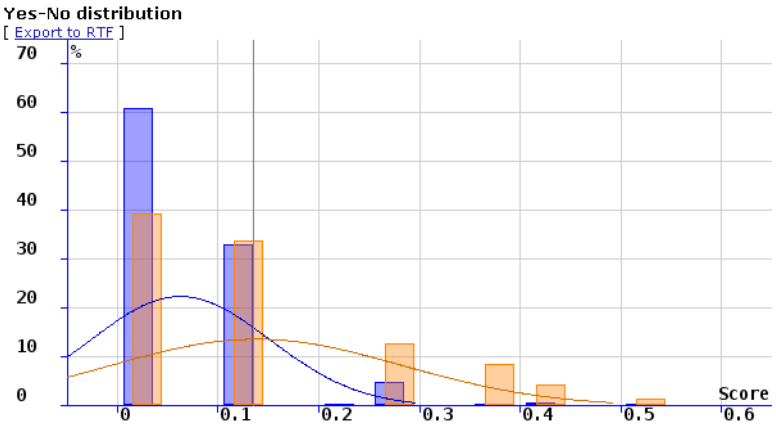


Figure 6.8 "Yes-No" distribution

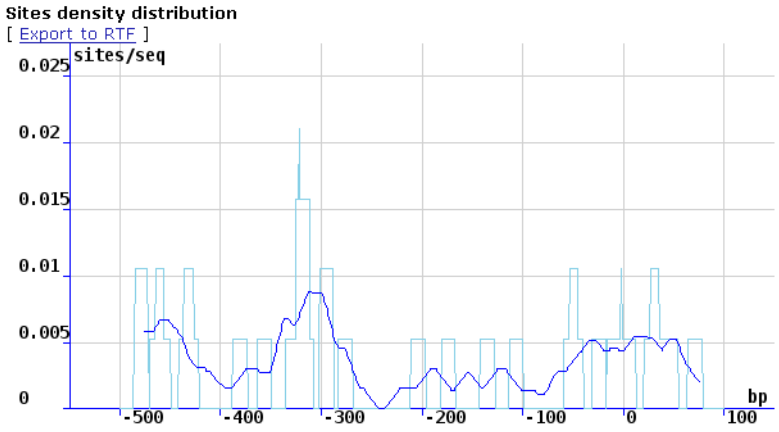


NOTE

The CMA overall cut-off is shown in both plots by a vertical line.

The "Sites density" distribution graph shows the number of sites lying in a specific sequence position (relative to TSS) divided by total number of sequences. For your convenience a smoothed graph is also displayed whith an average site density of 50 bp.

Figure 6.9 "Sites density" distribution



By pressing the [Show sites on sequences > >] button shown below the [+ Expression-score and Yes/No distribution] link, you are provided with information about matrix and model matches on promoters. A more detailed description of promoter extended model view will be given in the Section 6.2.

The "Model search result" menu appears when you select the CMA node in the project tree. From the dropdown menu you can select the "Get matrices of the model as gene set" link to compile a new data set from the factors linked to PWMs of the model. The link "Save model" adds a node for the model to the project tree. A model saved this way can be used for promoter classification as described in Section 6.6. The "Edit" link launches model editor described in Section 6.5.

As mentioned before, the best model with the highest fitness score is shown as the CMA output. By clicking on the [View model list] link above the model description, you will see the list of several models found by CMA with information about one model per row. The "Promoter model" column describes the PWM composition of the models with a concise syntax. Each CM component of a model is identified by a module number, e.g. M1, and described by single matrices and matrix pairs enclosed in squared brackets. For instance, the term

[V$EN1_01|V$SOX9_B1|V$NFAT_Q6|V$RP58_01>*V$CHX10_01|V$GCM_Q2**V$CIZ_01]

describes a module consisting of three single matrices and two matrix pairs. The single matrices are V$EN1_01, V$SOX9_B1 and V$NFAT_Q6. There is one pair of the matrices V$RP58_01 and V$CHX10_01 in which V$RP58_01 can only occur in direct orientation (>) and V$CHX10_01 can occur in both orientations (*) as well as one pair of the matrices V$GCM_Q2 and V$CIZ_01 which can both occur in either orientation (*). The CM definition part is followed by the definition of the complete promoter model by its modules (PM=...). Since all models in Figure 6.10 contain only one CM component, this definition is always PM=M1.

The "Model name" column assigns a unique name to each model. Columns "R", "T", "E", "N" and "P" provide values achieved in the corresponding fitness components described in Section 6.8.2. The total fitness of a model is given in the "Fitness" column. By revealing hidden columns you can add into each row a set of matrices comprising a model, gene symbols corresponding to matrices and hypergeometric distribution based p-value. Entries of the "Promoter model" column link to comprehensive reports about the corresponding model described before.

You can use the "Model search result" menu to operate with models marked in the checkbox column. The "Get matrices of the model as gene set" link will compile a new data set from the factors linked to PWMs of the selected models. The "Save model" link will add a separate node for each model to the project tree.

Figure 6.10 CMA output table

[+] **Model prediction results:** 3-5s 1mod 6gr HIF

Filter (filter bar): none (total 51 rows) Rows per page: 500

Export: Plain text | XLS | RTF

Mark: All (51) | None | Invert

Promoter model	Model name	R	T	E	N	P	Fitness
M1=[V$HIF1_Q5|V$MAZR_01|V$MEIS1AHOXA9_01|V$MMEF2_Q6|V$VDR_Q3] M2=[V$CBF_02|V$GATA3_02|V$HIF1_Q5|V$P53_DECAMER_Q2] M3=[V$HFH3_01|V$HIF1_Q5|V$MAZR_01|V$MMEF2_Q6|V$VMYB_02] PM=(M1|M2|M3)	PM1	nan	0.699202	0.767814	0.839957	0.305628	0.733508
M1=[V$HFH3_01|V$HIF1_Q5|V$MAZR_01|V$MMEF2_Q6|V$VMYB_02] M2=[V$CBF_02|V$GATA3_02|V$HIF1_Q5|V$P53_DECAMER_Q2] M3=[V$HIF1_Q5|V$MAZR_01|V$MEIS1AHOXA9_01|V$MMEF2_Q6|V$VDR_Q3] PM=(M1|M2|M3)	PM2	nan	0.693391	0.76681	0.604772	0.304208	0.7301
M1=[V$HFH3_01|V$HIF1_Q5|V$MAZR_01|V$MMEF2_Q6|V$VMYB_02] M2=[V$HIF1_Q5|V$MAZR_01|V$MEIS1AHOXA9_01|V$MMEF2_Q6|V$VDR_Q3] M3=[V$CBF_02|V$HIF1_Q5|V$LFA1_Q6|V$MYCMAX_02|V$SPY_02] PM=(M1|M2|M3)	PM3	nan	0.69724	0.761675	0.724287	0.291783	0.729458
M1=[V$CBF_02|V$GATA3_02|V$HIF1_Q5|V$P53_DECAMER_Q2] M2=[V$HFH3_01|V$HIF1_Q5|V$MAZR_01|V$MMEF2_Q6|V$VMYB_02] M3=[V$HIF1_Q5|V$MAZR_01|V$MEIS1AHOXA9_01|V$MMEF2_Q6|V$VDR_Q3] PM=(M1|M2|M3)	PM4	nan	0.692148	0.765806	0.604772	0.30374	0.728977

6.2 Composite modules on promoters

By pressing the [Show sites on sequences > >] button shown below the "Yes-No" distribution" chart, you will be provided with information about matrix and model matches on promoters. This button will be changed to [<< Hide sites on sequences] which will return to the view without promoters. The switch between main and background set View: **yes set (...)** | background set (...) also appears in extended view. The promoters table below the graphical description of the model is very similar to the one found in the MATCH report described in Section 5.2. The PWM color legend is followed by a table with one row for each promoter. The "Match display" column contains graphical presentations of matches along the promoter region with CMA model matches as grey boxes. Single TRANSFAC matrix matches are displayed as single colored arrows. Pairs of sites are displayed as two linked arrows. In the PWM color legend, pairs are displayed as two arrows in the same cell. For each matrix or pair, impact value is calculated which roughly estimates the contribution of a given matrix or pair to the final fitness value of whole CM using the following formulae:

$$I'(m_i) = F(M) - F(M \setminus m_i)$$

$$I(m_i) = \frac{I'(m_i)}{\sum_j I'(m_j)}$$

Where m_i is i-th component (matrix or pair), $I(m_i)$ is corresponding impact value, M is the whole CM, $M \setminus m_i$ is CM without component m_i and F is the fitness value. Thus if the impact value is close to 1, then the corresponding component contributes greatly to the overall fitness. Note that sometimes you may observe a negative impact value, which means removing the corresponding component from CM would increase the score. Usually this means that CMA did not run long enough to produce a good model.

Nucleotide positions are given relative to the TSS which is indicated by a kinked arrow. The "Sequence score" column shows the model score on the promoter.

The table also contains columns with the TRANSPro ID of the sequence, corresponding gene name, description of the corresponding gene, chromosomal location of the promoter. Columns linked from parent sets are available through the column hiding mechanism.

The table enables users to select individual promoters in the checkbox column. In the "Model search result" menu you may save selected promoters as a gene set by clicking on the "Get selected promoters as set" menu link. By clicking on the "Show text report on selected promoters" menu link you obtain a detailed text report of the selected promoters. Clicking on a picture of a promoter in "Picture" area provides detailed view of the promoter with site viewer (see Section 5.2.4).

Similar to the Match result additional columns were added to represent score contribution of each model component (matrix or pair).

If there are no sites for given component on some promoters, then a gray zero is displayed. When sequence score is calculated, for each module the sum of the component scores within it is calculated, then for each group the minimal score is selected (fuzzy AND operator) and finally maximal group score is selected (fuzzy OR operator). If a repressing group is present, its score is subtracted from 1 (fuzzy NOT operator).

Figure 6.11 Promoter description in the model report table

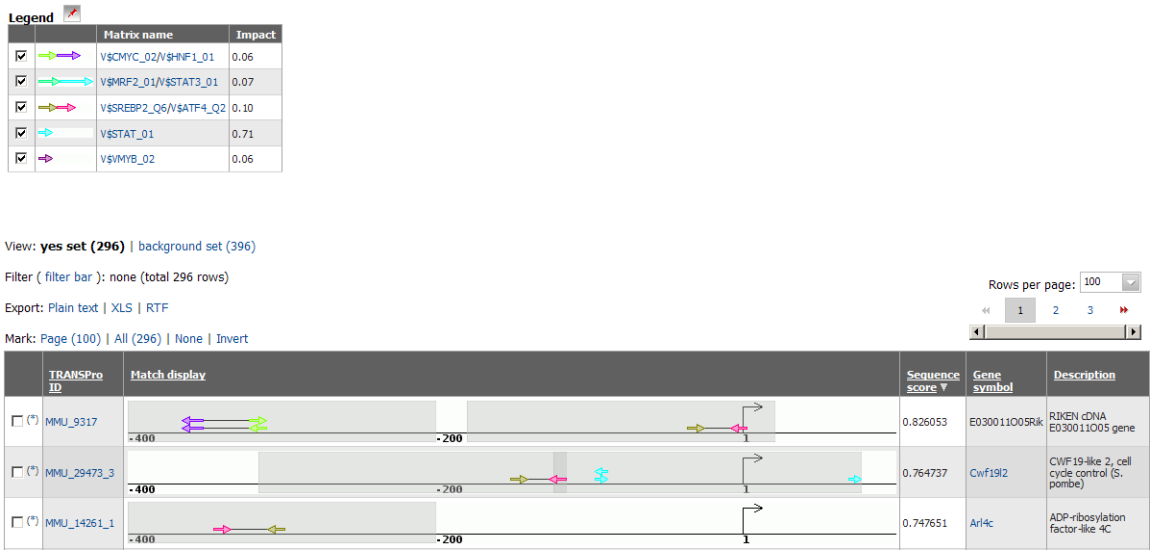


Figure 6.12 Columns displaying scores of individual matrices/pairs

	Sequence score	Gene symbol	Description	Affy ID	TSS	V$MRF2_01>>V$STAT3_01	V$STAT_01	V$VMYB_02	V$CMYC_02**V$HNF1_01	V$ATF4_Q2>*V$SREBP2_Q6
	0.826053	E030011O05Rik	RIKEN cDNA E030011O05 gene	1443677_at	Chr.9 96653146 (–)	0	0	0	0.09488	0.826053
	0.764737	Cwf19l2	CWF19-like 2, cell cycle control (S. pombe)	1453688_at	Chr.9 3478219 (+)	0	0.344875	0	0	0.764737
	0.747651	Arl4c	ADP-ribosylation factor-like 4C	1454788_at	Chr.1 90598488 (–)	0	0	0	0	0.747651
	0.744336	Fcho2	FCH domain only 2	1442453_at	Chr.13 99585252 (–)	0	0.101	0	0	0.744336
	0.74378	Vti1a	vesicle transport through interaction with t-SNAREs homolog 1A (yeast)	1439876_at	Chr.19 55390458 (+)	0	0	0.124	0	0.74378

6.3 Predefined parameters of CMA

The parameters set saved in the CMA advanced dialog window appears in the project tree inside the ”CMA” subfolder of the ”Preset” folder. When you select a preset node in the tree, all parameters are displayed as a set of tables similar to ones in the dialog window.

ExPlain provides two types of predefined parameters. System presets are loaded at the first start of the ExPlain application and cannot be removed. The example below shows one of the system presets. The second type of presets can be created through CMA advanced dialog window. When you launch the CMA dialog and follow the >> Next option once or twice, you see a checkbox with the field specifying the name of new preset to be saved. Mark this checkbox and start the CMA, then preset will be saved to the project tree. User-created presets can be managed as other items in the tree.

6.4 Composite element models

The ExPlain application provides several kinds of models that can be used to classify promoters of some data sets.

Composite Element models for CMA which were compiled on the basis of TRANSCompel composite elements [1] can be found in the folder ”Composite Elements” in the project tree. As system preloaded models, they cannot be removed or modified.

Any model when selected is displayed graphically. The extended description is available after clicking on [Verbose display] link above the model. The [TransCompel model link] button links to the TRANSCompel model description. The example below shows an extended view of the system model named CEPB/NFkappaB.

Models saved from the editor (see Section 6.5) also have a graphical representation and can be viewed in a simple and extended mode.

Figure 6.13 CMA preset parameters

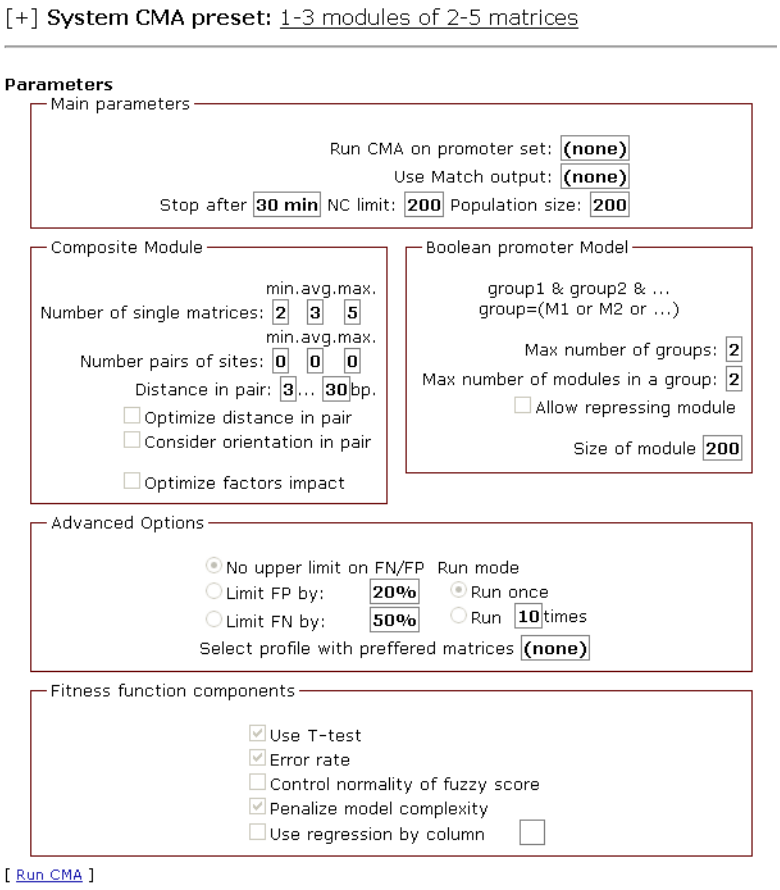

[+] System CMA preset: 1-3 modules of 2-5 matrices
Parameters
Main parameters
Run CMA on promoter set: (none)
Use Match output: (none)
Stop after 30 min NC limit: 200 Population size: 200
Composite Module
min.avg.max.
Number of single matrices: 2 3 5
min.avg.max.
Number pairs of sites: 0 0 0
Distance in pair: 3... 30bp.
Optimize distance in pair
Consider orientation in pair
Optimize factors impact
Boolean promoter Model
group1 & group2 & ...
group=(M1 or M2 or ...)
Max number of groups: 2
Max number of modules in a group: 2
Allow repressing module
Size of module 200
Advanced Options
No upper limit on FN/FP
Limit FP by: 20%
Limit FN by: 50%
Run mode
Run once
Run 10 times
Select profile with preffered matrices (none)
Fitness function components
Use T-test
Error rate
Control normality of fuzzy score
Penalize model complexity
Use regression by column
[Run CMA]

Figure 6.14 System model

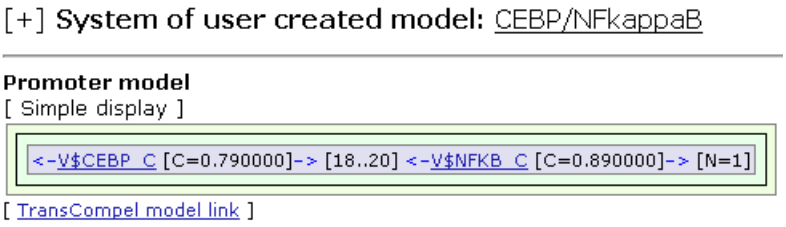
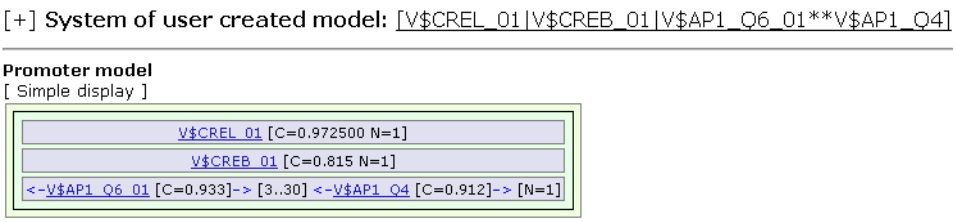


Figure 6.15 User created model



Both system and user-created models have the ”Model” menu, where you can edit any model using the ”Edit” link.

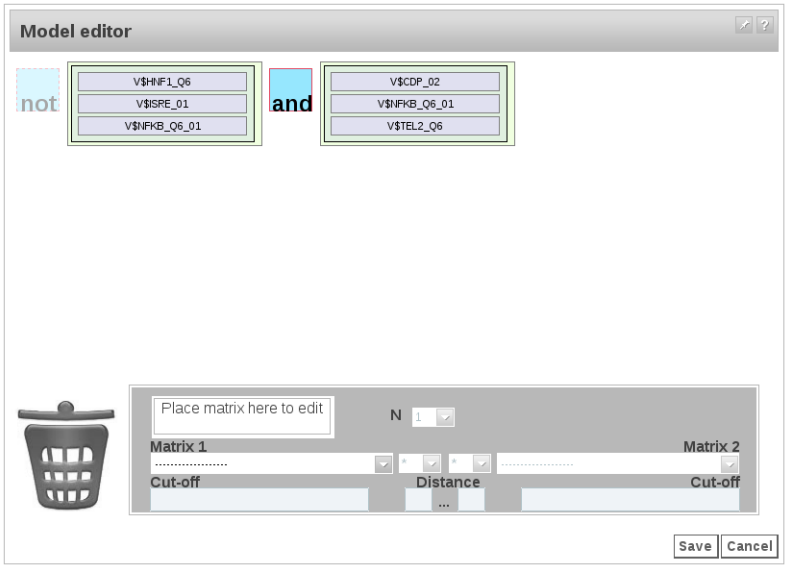
6.5 Model editor

ExPlain gives you the opportunity to create a boolean promoter model with any parameters and any structure. The boolean model can be generally represented as union group1 & group2 &. . . , where each

group is combined out of modules: group = (M1 or M2 or . . .), modules can include single matrices and matrix pairs. A repressing module is included in the model with the logical operator NOT.

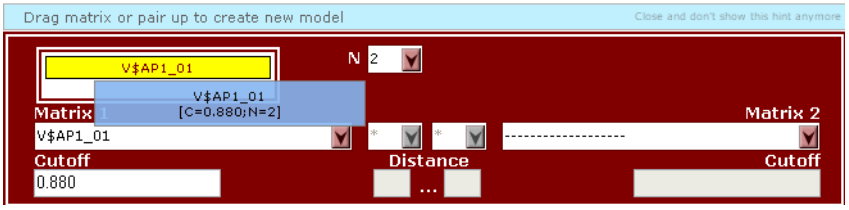
There are two ways to run the model editor in ExPlain. After clicking on the "New model" link in the "File" menu, the model editor will be launched with an empty input window (see the figure below). If you are choosing the "Edit" option with a model (calculated by CMA or acquired by any other way) as a current active node, the current model will be opened in the model editor.

Figure 6.16 Model editor window



To add a single PWM to a new model, first select it in the drop-down box "Matrix 1". We chose V$AP1_01 as an example. The selected matrix appears in the box above. Edit the cut-off value in the "Cut-off" field, and select the N - number of matrix matches in the module. If you move your mouse pointer on top of the matrix, you will see a floating blue box with the matrix parameters inside. When you are ready, drag the created matrix from the editor box up to create a new model.

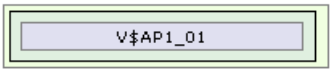
Figure 6.17 Adding a single PWM



To add a pair of PWM to a new model, select matrices in "Matrix 1" and "Matrix 2" drop-down boxes, edit the cut-off values, choose the orientation of matrices in the pair, the spacer range, and the number of pair matches in the module. Drag the created pair from the editor box up to add it to the model.

The boolean structure of the model is represented by colored boxes. A single matrix or matrix pair appears inside a gray box, and a single matrix or several matrices and pairs comprising a group will appear within a light green group rectangle. An entire model will appear in a yellow box, consisting of groups containing matrices or matrix pairs.

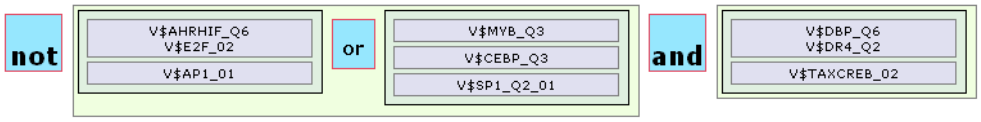
quation 6.5.1



To add matrices or pairs to the model as a part of the module, group, or a model itself drag the gray rectangle with corresponding data inside the proper coloured rectangle. Click the "not" button to make

a module repressing. If you want to remove some elements, drag them to the trash can at the bottom-left corner. It can be difficult for the first time to adjust all the elements, but after some practice it will be much easier.

Figure 6.18 An example of a model created in the editor



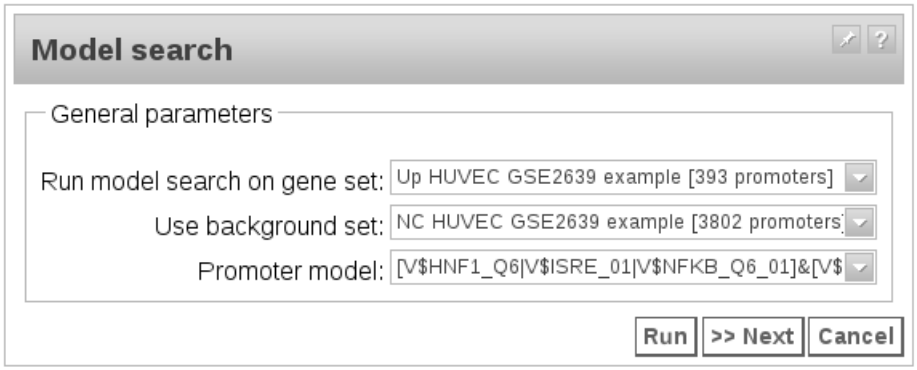
When you are satisfied with your model, press the Save. If you created a new model, a new model node will be added to the "Composite elements" folder in the process tree. If you were editing an already existing model, the new node will appear under the original model.

6.6 Classifying promoters

You can use CM models to classify promoters of other datasets. This functionality is made available through the "Scanning pre-defined CMs" link of the "Analyze" menu.

In the dialog window it is necessary to choose a gene set and a composite model. If the current selected node is a CM saved from CMA or one of those described in Section 6.4, it will be selected in "Promoter model" field of the dialog window. If the current node is a gene set, then it will automatically be selected as the main gene set.

Figure 6.19 CM scanning dialog window



You can launch the search by pressing the Run button.

By pressing the >> Next you will be able to set up more parameters in the advanced dialog window. By default the parameters take values assigned to the selected model. The length of the promoter sequences, type of promoters to use, size of module and fitness function parameters from CMA (see Section 6.8.2) can be adjusted here.

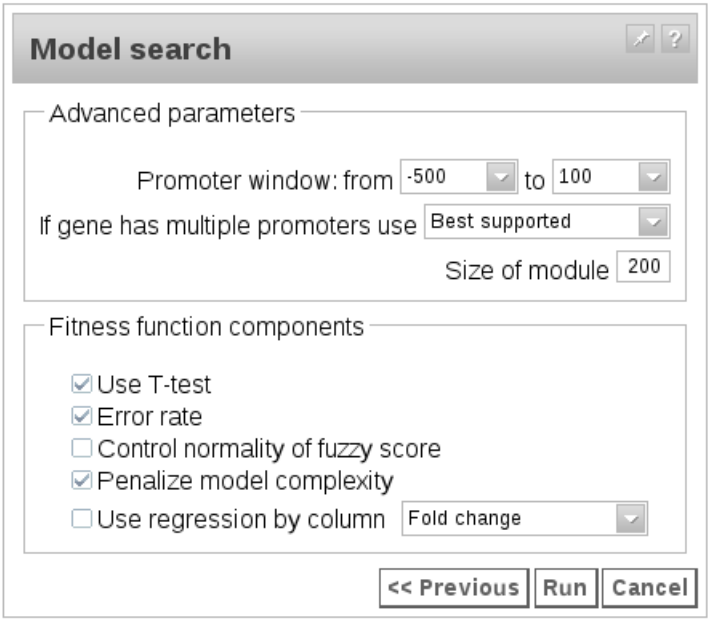
The CM scanning output looks exactly the same as the one described in Section 6.1.3, including the graphical description, calculation performance, statistical distribution plots and promoter table.

The "Match display" column contains a graphical representations of matches along the promoter region, with CMA model matches as gray boxes and TRANSFAC matrix matches as colored arrows. The "Sequence score" column represents the calculation score of the sequence. This column is automatically added to the main set. If you select the main set node in the project tree, you will see the scanning results column named by CM.

6.7 Obtaining interactions between Transcription factors and their target genes

Once you have a list of putative targets of a promoter model obtained by CMA, you can generate a list of interactions between the transcription factors present in the model and the target genes where their

Figure 6.20 CM scanning advanced dialog window



sites are located.

This function allows you to include the regulatory interactions between the transcription factors found in a CMA run, and their target genes. This interaction profile can then be used to enrich Key Node analysis. This function can be launched by clicking on the icon, visible when standing on a CMA search result node. In the dialog window it is necessary to choose a gene set and a composite model. If the current selected node is a CM saved from CMA or one of those described in the Section 6.4, it will be selected in the "Promoter model" field of the dialog window. If the current node is a gene set, then it will automatically be selected as the main gene set.

You can launch the search by pressing the Run button.

The program has two parameters to set:

a)The result of searching a saved CMA model on a gene set

b)The percentage (%) of false positives to allow when setting the threshold (set to 10 for a relaxed threshold, and to 1 for a stringent one).

The result of this program is a table with pairs of molecules, corresponding to the TFs from the CMA model and the genes targeted by these TFs. This table can be added as "user interactions" when running a key node analysis.

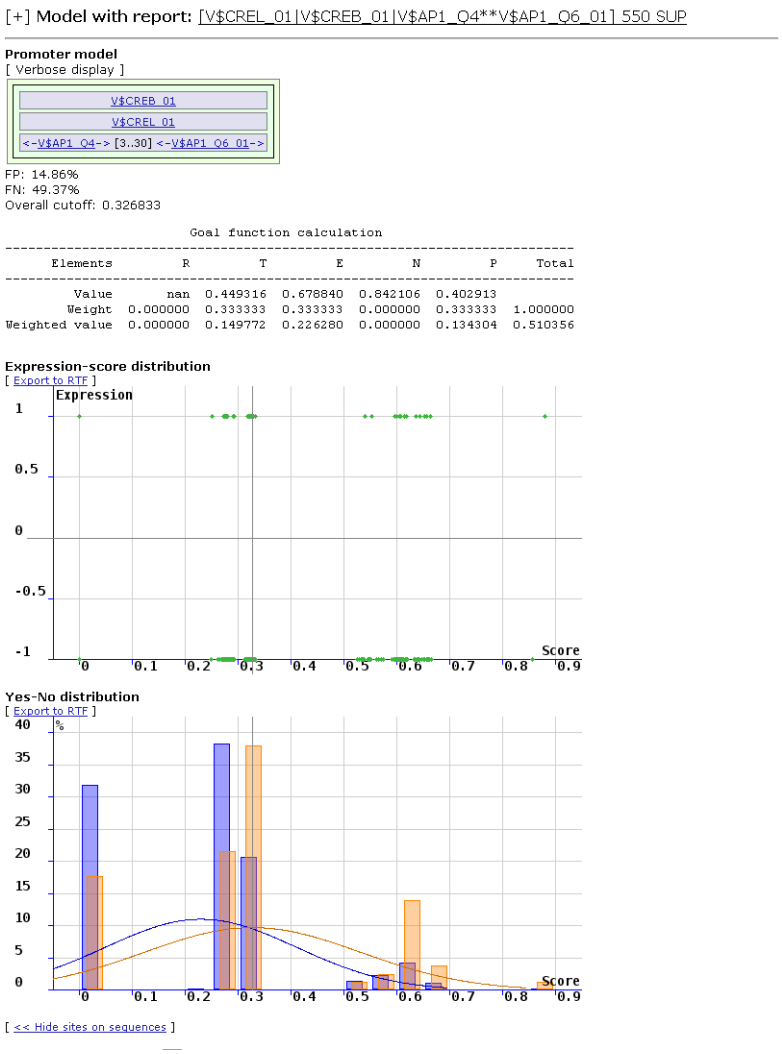
Algorithm.

The interaction profile generator reads the score of the CMA model in each target gene, in both the background set used to run CMA and the set of new targets which are the product of the CMA search. The percentage of false positives passed as a parameter is used to decide the exact CMA score that would identify the given percentage of genes in the control set as being regulated by the CMA model, and use this as a threshold. The genes in the control set are however not included in the interaction table. Setting a large percentage implies a low CMA score threshold and will produce an interaction table that will include many of the real target genes. The use of a smaller percentage will produce a higher score threshold, and an interaction table with fewer genes.

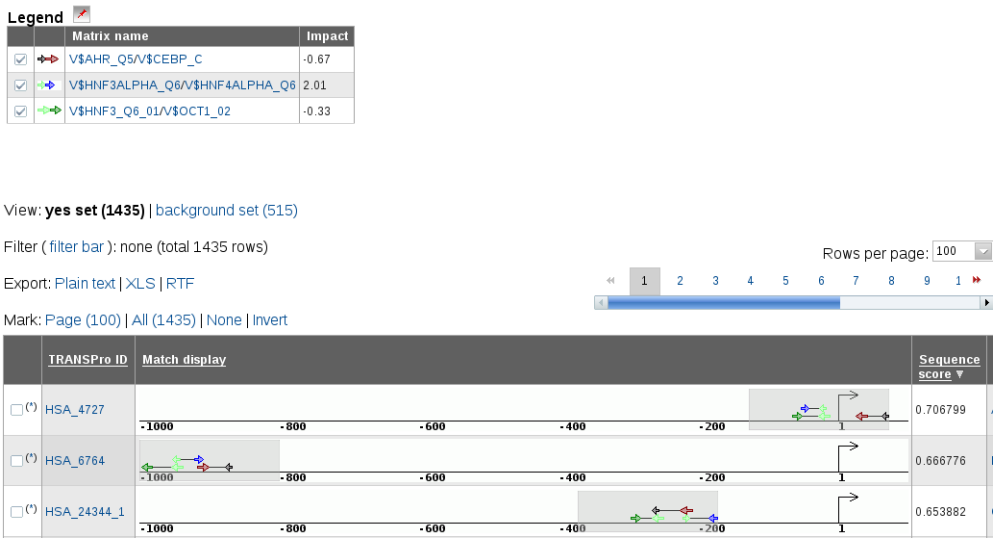
6.8 CMA - Composite Module Analyst – Background information

This section provides an overview of the Composite Module Analyst (CMA) software. CMA is a software tool that produces a model of the common regulatory regions in a promoter set, described by combinations of single transcription factor binding sites as well as their pairs in composite modules. A composite module is considered to be a region within a promoter in which specific combinations of

Figure 6.21 CM scanning output



(a)



(b)

binding sites are located in close proximity (e.g. 300 nt) to each other.

The software receives as input a set of positive promoters as well as a set of negative promoters

Figure 6.22 CM scanning score column

Gene Symbol	TSS	VAHR_Q5**VCEBP_C	V$HNF3ALPHA_Q6**V$HNF4ALPHA_Q6
ALDH1A1	Chr.9 74757786 (−)	0.246453	0.213256
LAMP2	Chr.X 119487177 (−)	0.240404	0.212808
CCNB1IP1	Chr.14 19871268 (−)	0.238733	0.194972
TP53TG1	Chr.7 86812741 (−)	0.24124	0.206963
ZNF330	Chr.4 142361520 (+)	0.250682	0.181715

Figure 6.23 Create interactions dialog window

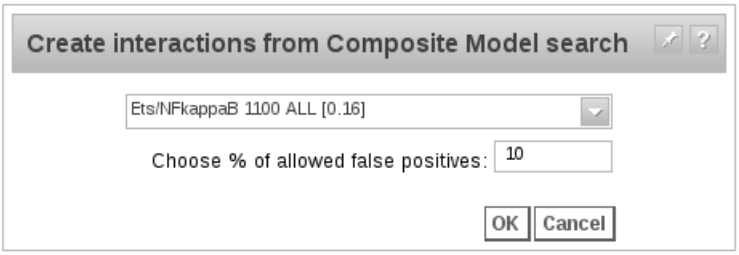


Figure 6.24 CM interaction output

[+] **Interaction profile:** reactionset

Filter (filter bar): none (total 36 rows)

Rows per page: 100

Export: Plain text | XLS | RTF

Mark: All (36) | None | Invert

From molecule name	From ID	To molecule name	To ID	Cost	ActRep
AChE	MO000068336	eIF-4E	MO000020752	2	1
alpha2-HSG	MO000044587	alpha2A-AR	MO000018309	1	1
CP2a	MO000054371	CP2 isoform 2	MO000036897	0	1
CP2a	MO000054371	CP2-xbb2	MO000036897	0	1
CP2a	MO000054371	CP2a	MO000036897	0	1
CP2a	MO000054371	LBP-1d	MO000036897	0	1
DeltaNp73beta	MO000044149	TACE	MO000017665	2	1

together with MATCH binding site predictions. The first set may correspond to genes found to be differentially expressed during an experiment whereas the second may be a set of genes one does not expect to respond to the same signal as the target genes, e.g. genes which did not exhibit differential expression during the same experiment. CMA then attempts to find a combinatorial motif model that best discriminates between positive and negative promoters.

6.8.1 CMA promoter models

The main component of a CMA model is the composite module (CM). Each composite module is defined by 3 parameters (ϕ,M,ψ), where ϕ is the set of transcription factors regulating a promoter, M is a set of PWMs used to predict binding sites for the factors of ϕ, and ψ is a set of parameters defining the structure of the CM, such as the window size, the number of single motifs, the number of pairs of motifs and score cut-offs for pairs and singletons. CMA promoter models can combine multiple CMs in groups and also be composed of more than one group.

Promoter sequences are classified by sliding a window of the size of a CM along the sequence and scoring each according to the CM parameters to finally yield a normalized score over all window positions. The normalized CM score is transformed to 0 (negative) or 1 (positive) according to a threshold

defined for each CM of the model. These binary scores are then combined by a function of Boolean operators which classifies a promoter as positive if the overall output is 1.

6.8.2 Model construction

CMA attempts to find a model that optimally discriminates between promoters of the positive and the negative set. In the absence of an analytic method to compute optimal parameters, this task is assessed by stochastic optimization with a Genetic Algorithm (GA).

Briefly, a GA works iteratively (i.e. over several discrete generations) on a population of solutions. In the case of CMA, the solutions are promoter models defined by their parameters. Typically, a population of a new generation is created by selecting solutions (individuals) of a population according to their performance (fitness) and introducing variation (mutations), whereas the current best solution is additionally taken to the new population directly without modification. Hence, a single iteration consists of creating a new population based on previous results and testing the performance of each solution. The initial population may be either created with randomly constructed solutions or roughly estimated parameters. As any stochastic optimization method, GAs require an externally defined termination criterion, e.g. the simplest one might be to optimize for a certain number of iterations or time interval.

CMA implements a five-component fitness function to evaluate the performance of its models. The fitness function components are described below. Each component is considered with a specifiable weight to obtain the fitness for a model.

CMA FITNESS FUNCTION COMPONENTS

R The R component measures how well CM scores fit the expression values (integers resembling categories, e.g. NC $\rightarrow$ 0 or continuous numbers, e.g. fold change) by linear regression.

T The T component assesses the statistical significance of the difference between the distributions of fuzzy scores derived from the Boolean function for each the two promoter sets by the t-test.

E The E component controls the error rate of a model. Here, false positive and false negative errors are considered and derived from classification performance on the two promoter sets.

N The N component controls the normality of fuzzy scores also used in the T component, i.e. their resemblance of a normal distribution. This parameter is included to prevent over-fitting.

P The P component penalizes model complexity according to the number of CM units and the matrices and pairs each CM contains. Therefore, it also safeguards against over-fitting.

For a full description of the CMA algorithm please see: Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. Waleev T, Shtokalo D, Konovalova T, Voss N, Cheremushkin E, Stegmaier P, Kel-Margoulis O, Wingender E, Kel A. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W541-5. PMID: 16845066

Chapter 7

Molecular networks analysis

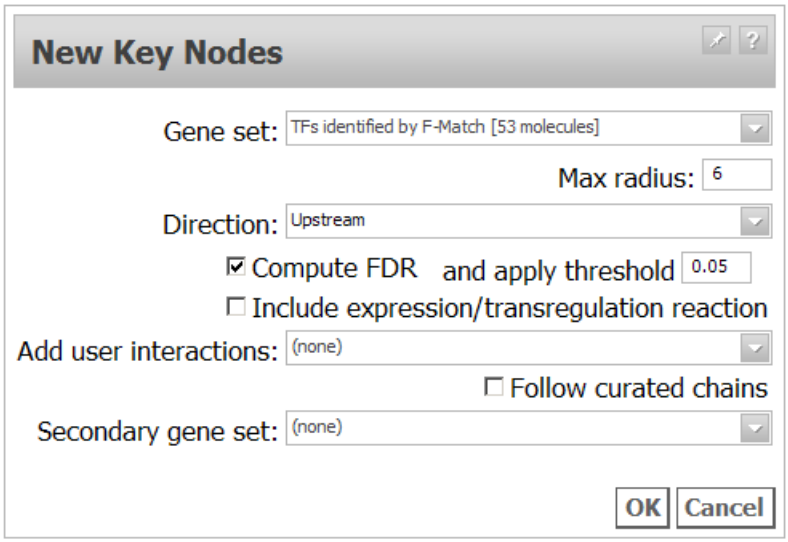
ExPlain provides information about signal transduction networks contained in the BKL database.

Using the search for ”Key nodes” option of ExPlain, you can search for common signaling molecules (key nodes) in the network vicinity of your gene set. The ”Network clusters” tab can be used to identify subnetworks containing genes coherently connected by signal transduction reactions.

7.1 Network key node analysis

7.1.1 Key nodes dialog window

Figure 7.1 Search for key nodes dialog window



To start a key node analysis open the key nodes dialog by clicking on the ”Key nodes” link in the ”Network analysis” section of the ”Analyze” menu. First you should select a gene set you want to analyze in the ”Gene set” field. By default the current tree node is selected as input set. The dialog provides several further input options that can be used to specify the conditions for the analysis.

Max radius The maximal search distance threshold defines the number of steps from each input molecule that are considered by the algorithm.

FDR You have the option to compute the False Discovery Rate and to filter the results by a FDR-threshold. Only key nodes with a FDR value below the specified threshold will be shown in the result list.

Expression/transregulation reaction If gene regulation and transregulation reactions shall be considered in the search for key node molecules, use the "Include expression/transregulation reaction" option. If this option is not checked all reactions including gene regulation / transregulation events will be excluded from the analyzed networks.

User interactions In the "Add user defined interactions" list, you can select any uploaded interaction profile (see Section 7.4 for more details). The Algorithm will consider the corresponding molecule interactions during the key node search.

Follow curated chains If annotated pathways and reaction chains are to be considered (marked "Follow curated chains" checkbox), the persistence reward controls the integration of reliable annotated pathways and reaction chains during the analysis.

Secondary set You may choose a secondary gene set to run a two-set network analysis in the "Use secondary set" parameters block. The set of secondary genes is then included by the algorithm, and reaction paths that go through the elements of this gene set are given higher scores.

The key nodes searching algorithm can be used to identify common signaling molecules in the network vicinity of genes from your input list.

The underlying application searches the network in a specified range starting from each input molecule in order to find the most proximal molecule that is connected to a maximal number of input molecules. This is achieved by scoring each node that was visited on the path from any input molecule.

Since the resulting score may be determined by a generally high level of connectivity of some molecule, the total number of connections reaching every node is also taken into account by the algorithm and penalized, in order to acquire a preference for molecules that are specific for your input genes.

The application is also capable of taking a secondary gene set into account and finding reaction chains that use the molecules described in this set.

Fully annotated pathways and reaction chains can be included in the search. Path predictions and expert annotation are dynamically combined and can be balanced by the user.

7.1.2 Results of key node analysis

We describe the output of the "Key Nodes" analysis by an example computed with a set of transcription factors obtained in a F-Match analysis for up-regulated genes from the HUVEC GSE2639 example (The way in which the set of transcription factors used as input for the key node analysis was derived is shown in Section 5.1).

Upstream reactions were searched with a distance threshold of 6, without including expression / transregulation reactions and "follow curated chains" options. The results were filtered by a false discovery rate threshold of 0.05.

Figure 7.2 shows the output frame of the network node created by the application. Each row of the output table contains information about one key node. The molecule name and classification are given in respective columns. The "#Hits in network" values represent the number of input molecules that are connected with the key molecule within the required distance. The "Hits list" column contains the names of the input molecules that are connected to the described key node. Furthermore there are three columns with the Score, Z-score, and FDR values calculated for each key node, which allow an evaluation of the key node reliability (see below).

Some additional columns ("TRANSPATH" id's for key molecules and hits, molecule type description, distance, and "#non-relevant reachable nodes") are hidden. In the "#non-relevant reachable nodes" column the number of molecules that could be reached, but were not present in the input list is displayed.

All key nodes (either molecule or gene) in the analysis results list will be ordered by default based on their significance score. This score reflects the relation of connected relevant nodes (i.e. the nodes that correspond to the molecule/gene list from your initial query) to nonrelevant nodes (i.e. molecules/genes that can be reached from the key node but are not in the initial list).

Details about the score calculation can be found in the "key node algorithm" section below. The false discory rate allows an evaluation of the probability of finding the respective key node at the same or a

Figure 7.2 Output frame of a key node analysis

	Molecule name	Molecule classification	#Hits in network	Score ▼	Z-Score	FDR
Filter »	(none)	(none)	(none)	(none)	(none)	< 0.05
☐ (*)	SHP1-isoform1	enzymes; hydrolases EC 3; hydrolases EC 3.1; phosphoric monoester hydrolases EC 3.1.3; protein phosphatases; protein tyrosine phosphatases; SHPs; SHP-1; SHP-1(h)	31	4.83637	1.14539	0
☐ (*)	SHP1-isoform2	enzymes; hydrolases EC 3; hydrolases EC 3.1; phosphoric monoester hydrolases EC 3.1.3; protein phosphatases; protein tyrosine phosphatases; SHPs; SHP-1; SHP-1(h)	31	4.83637	1.14539	0
☐ (*)	Caspase-6	enzymes; hydrolases EC 3; peptidases EC 3.4; cysteine endopeptidase EC 3.4.22; Caspases; Caspase-6	30	4.81581	1.14162	0.01
☐ (*)	5-HT-2A	membrane-transducing components; receptors; GPCR; rhodopsin like receptors (type A); amine receptors; serotonin receptors; 5-HT-2	27	4.808	1.95191	0.012

better rank position by random chance. Key nodes at the top of the result list that have small FDR values are thus very reliable, whereas key nodes with large FDR value have a high likelihood of appearing also at a significant position in any other analysis, and are therefore not very specific. The Z-Score serves as an additional measure of the key node reliability. Key nodes with a Z-Score above 1 can be regarded as statistically significant.

When the key node result is displayed, the custom menu "Network search result" appears in the menu bar. Corresponding icons are shown in the right side of the toolbar. You can obtain a combined presentation of several networks by marking key nodes in the checkbox column and pressing the "Networks - Visualize networks from selected rows" menu item or the toolbar button. There are several ways to derive subsets from rows marked in the output table:

The "Hits - Get hits from selected rows as gene set" command creates a subset from the input molecules connected to a key node.

Key node molecules themselves can be exported with the "Molecules - Get key molecules from selected rows as gene set" option.

Clicking the "Networks - All nodes between hits and key molecules from selected rows" link results in a subset containing all nodes from the whole network.

Figure 7.3 shows a part of the network of the top key node (SHP1). The key node is shown in pink, input molecules are indicated in blue, and nodes are displayed between in brown. Reaction types are represented by colored arrows and colored squares. A red square indicates inhibition, whereas a green square stands for activation. Gray arrows indicate semantic associations.

Read Section 7.3 for more information about the flash-based network visualization tool.

7.1.3 Summary of key node results

The results of several key node analyses can be combined in a summary table using the option "Keynodes summary" from the "Merge and summarize" section of the "Data" menu. In the create summary set dialog window (Figure 7.4), the key node results to be compared can be selected along with the

Figure 7.3 Network visualization

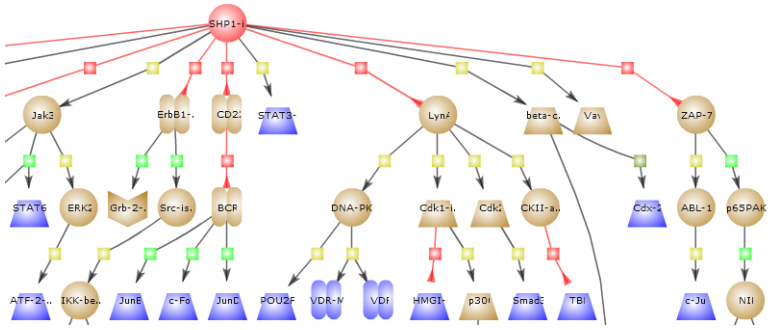
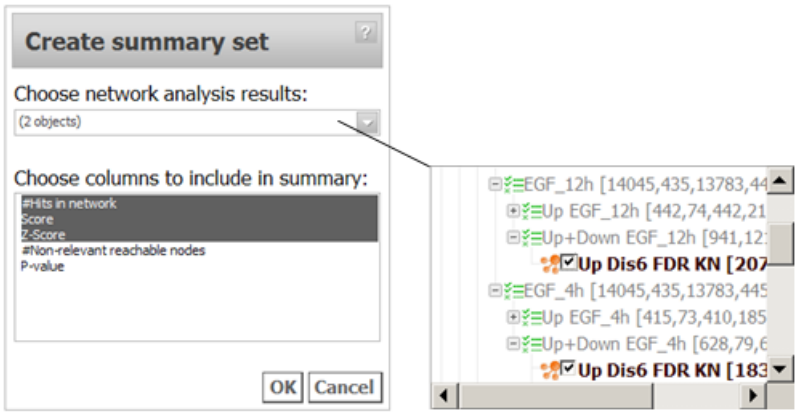


Figure 7.4 Create network summary dialog



columns that shall appear in the summary table. When the OK button is pressed, the process of summary creation will be launched.

In the summary output (Figure 7.5) the key node analyses results selected for summary display are shown in one table. It is now possible to directly compare the results of different key node searches, and to extract interesting molecules/genes via the option "Key molecules - Get key molecules from selected rows as gene set" of the summary specific menu.

Figure 7.5 Key node summary result

	Up Dis6 FDR KN (Up+Down EGF_12h)			Up Dis6 FDR KN (Up+Down EGF_4h)		
Molecule name	#Hits in network	Score	Z-Score	#Hits in network	Score	Z-Score
Lyn; Fyn; Blk{p}	14	2.59054	1.66514	14	1.60848	0.69456
A-Raf	12	2.24245	2.57663	0	0	0
Fyn{p}	14	2.17857	1.62622	14	1.60778	0.684806
Efs	14	2.17092	1.60809	14	1.60079	0.632798
integrins	15	2.13528	1.13169	16	1.8182	0.750911
IL-7R	14	2.12138	1.38576	16	1.841	1.03525
beta3-integrin	15	2.02969	1.19748	16	1.77834	0.588919
Abl	16	2.02135	0.319107	18	2.01898	0.325733
CD45	14	2.00433	1.31022	16	1.82912	0.986937
Csk	15	1.97145	0.688935	18	2.0558	0.547748
ERK2{p}	15	1.9636	1.56381	0	0	0
Src	16	1.95915	-0.122634	16	1.79937	-1.57795
CD19	15	1.95311	1.06172	17	1.82541	0.756598

7.1.4 Key node search algorithm

The search for signaling molecules (key nodes) in the network vicinity of a gene list can be performed based on only one gene list, or based on a primary gene list regarding an additional gene list as secondary set. Genes of a secondary set are incorporated such that the key node algorithm is pushed to go through the elements of this gene set. The network path is attracted by the secondary genes, resulting in longer paths being often cheaper than shorter paths if they include molecules from the secondary set. The algorithm is a feed-forward based approach which transforms the original weights of the network into new weights. The weights of the resulting network are such that the desired attraction power is reflected.

Score: The significance score, used for ranking the key node results, counts the hits relative to the respective logarithmized Volume V_i that was required to reach every hit.

quation 7.1.1 Score calculation

$$S = N_0 + \sum_{i=1}^{r} \frac{N_i - N_{i-1}}{1 + \log(V_i)}$$

Volume V_i = number of compounds reachable in total from the key node within distance i

Hits N_i = number of targets reached by key node $\mathbf{k}$ within distance i

With growing distance i the Volume V_i is growing too. The maximum distance is limited by r.

False-Discovery-Rate (FDR): Each individual key node gets a FDR-value assigned, which represents the probability to occupy the observed rank or higher ranks by random chance. It is estimated on-the-fly by random sampling. The ranking of the key nodes is defined by sorting them according to the score described above in descending order. All key nodes that have an observed rank lower than 200 get assigned 1.0 as FDR-value by definition, since their score is considered not to be sufficient. Molecules which do not have any hits get assigned the last rank since the score is zero in this case.

Z-Score: In addition to the FDR, each key node gets a Z-Score, which measures the deviation of the observed rank of the key node from the expected rank in random case, divided by the standard deviation.

quation 7.1.2 Z-Score calculation

$$Z = \frac{X - \mu}{\sigma}$$

In this formula, the rank distribution is assumed to comply with the normal distribution. Key nodes with Z greater than 1.0 are considered significant.

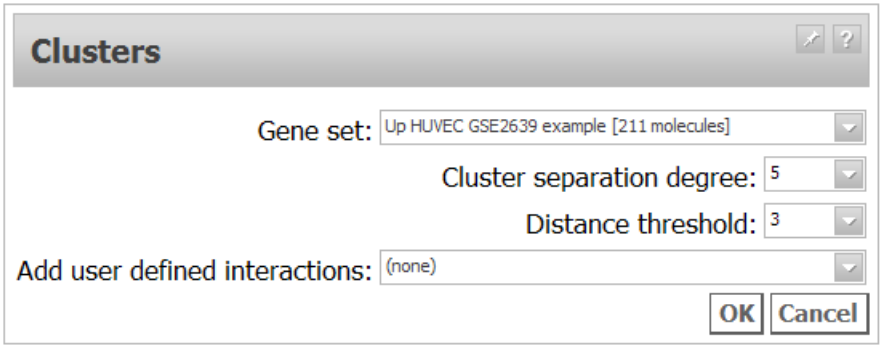
7.2 Network cluster analysis

7.2.1 Cluster dialog window

The ”Network clusters” analysis can be used to identify common subnetworks for molecules of a data set. The algorithm tries to connect each pair of molecules of the input set.

To start the analysis, open the ”Network clusters” dialog via the link in the ”Analyze” menu, and select the gene set you want to analyze in the ”Gene set” field. By default this field is set to the current tree node. Then select the parameters that will be used by the algorithm.

Figure 7.6 Search for clusters dialog window



Cluster separation The cluster separation degree influences the degree to which clusters are separated/divided and thus determines the cluster size. The higher the degree, the more edges are removed. The edges are assigned a betweenness value, and edges with high value are more likely to be removed. A low separation degree yields a single large cluster which is sometimes difficult to visualize. A high separation degree value can leave the input set unclustered. The size of the input list also influences this parameter: large inputs usually require higher separation degrees.

Distance threshold The maximal search distance threshold defines the number of steps from each input molecule that are considered in the cluster calculation. Only molecules that are linked by a number of steps smaller than the one specified by the distance threshold are connected by the algorithm.

User defined interactions In the "Add user defined interactions" list, you can select any uploaded interaction profile (see Section 7.4). The Algorithm will include these molecule interactions during the clusterization.

7.2.2 Network cluster analysis results

Figure 7.7 shows an example output of a cluster analysis conducted for the 300 most up-regulated genes from the HUVEC GSE2639 example (extracted in the same way as in Section 3.3.5) with cluster separation degree 5 and distance 3. The "#Hits in network" values represent the number of input molecules that are present in a certain subnetwork. Names and TRANSPATH accession numbers of input molecules are listed in "Molecule name" and "Molecule acc" columns respectively.

Figure 7.7 Network cluster output

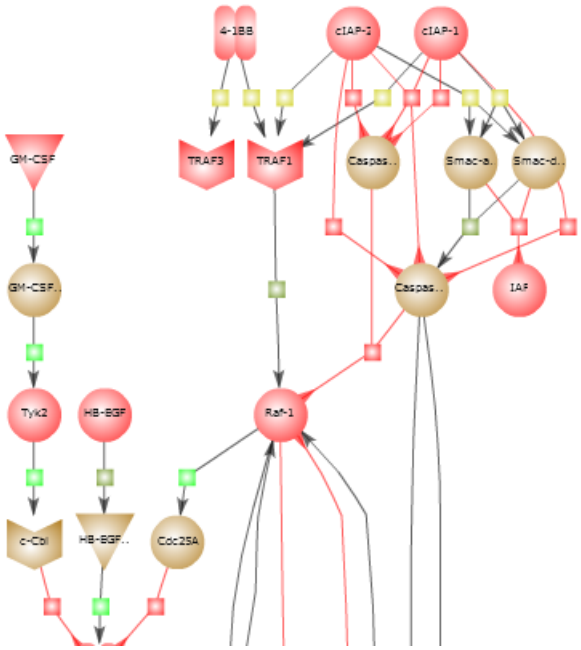
#Hits in network	Molecule name
17	4-1BB, Bcl-3, cIAP-1, cIAP-2, ErbB1-p170, GM-CSF, HB-EGF, IAP, IkappaB-alpha, inhibin beta A, MEK1, p62-isoform1, Raf-1, Smad3, TRAF1, TRAF3, Tyk2
5	beta-c, MKK3, SAB, STAT5A, TAB2-isoform1
3	IL-1alpha, IL-1beta, IL-1RI
2	A20, Naf1alpha
2	Apo2L, DR5
2	IL-15, IL15RA
2	Jagged1, Notch2

When a network cluster node is active, the custom menu "Clusters search result" , which provides specific actions with selected rows, appears in the menu bar. Corresponding icons appear at the right of the toolbar. You can obtain a representation of one or several clusters by marking key nodes in the checkbox column and pressing the "Clusters - Visualize selected clusters" menu item or toolbar button.

The "Hits - Get hits from selected rows as gene set" option creates a subset from the input molecules present in the selected subnetworks.

The network of the top cluster identified (cf. Figure 7.7) is displayed in Figure 7.8. In this visualization, pink nodes are the clustered input molecules.

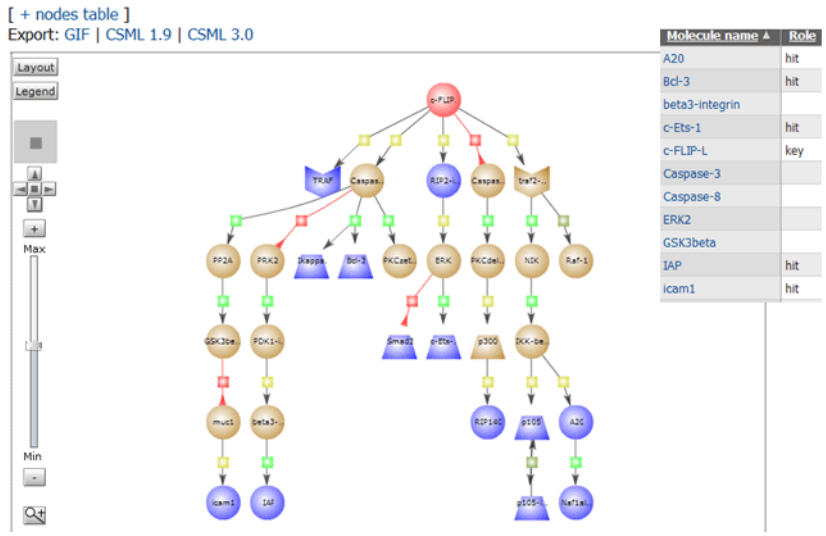
Figure 7.8 Section of a network cluster



7.3 Network visualization

A flash-based application provides you with tools to visualize and manipulate the network. The screen contains two parts: a navigation bar to the left, and a main canvas. You can change the scale, scroll the displayed network, view the information associated to each node or edit the network.

Figure 7.9 Flash-based network vizualization

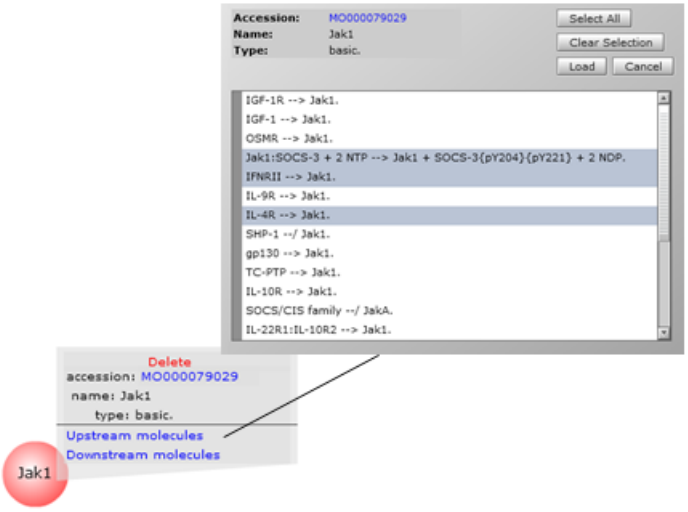


The network visualizer navigation bar consists of three navigation elements. The Navigation panel is a miniature scheme of the canvas, which shows the position and moves the visual window around big or zoomed-in networks. The panning buttons provide another tool to navigate stepwise within the

canvas. The zoom panel has a slider and buttons. The (+) button zooms in and the (-) button zooms out. There are several available export options shown above the network picture.

The area where the network is visualized is called the canvas. A network is a set of nodes which are connected to each other by edges. The nodes can be either molecules, genes or reactions. The node menu displays accession link, name and type of node and provides options to delete the node (”Delete”) or add some other sets of molecules to the network. By clicking on a node set, e.g. ”Upstream molecules”, you will get a list of the reactions which are upstream from the selected node.

Figure 7.10 Node options



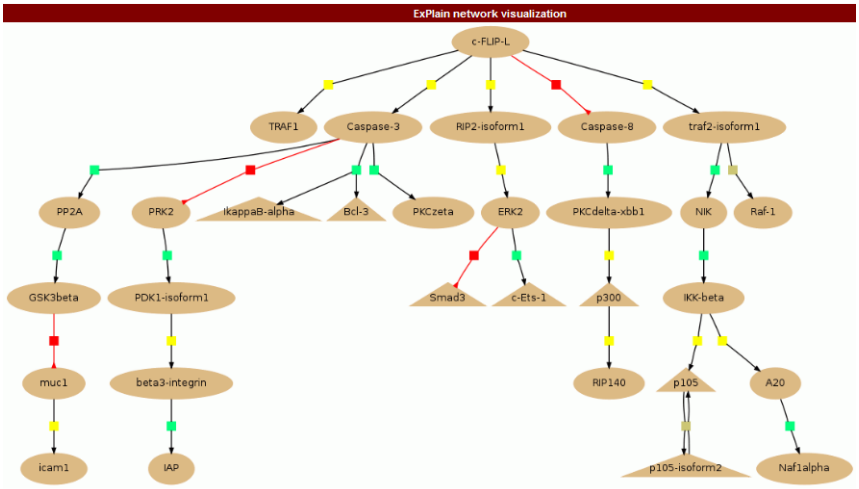
Colors and shapes of the network nodes distinguish molecule types and functions. Molecules are represented by ellipses, while two vertical ovals represent a receptor. Ligands have a triangle shape, and transcription factors a trapezium shape. Red molecules are key nodes, and blue ones represent end targets. Several nodes placed one over another denote complexes. All this information is available by clicking on the ”Legend” button.

The ”Layout” button refreshes the canvas, positioning the molecules hierarchically and adjusting the edges between them. This option should be used when you have done some changes to the network (like moving some molecules or edges or adding or removing some reactions) and want the network to look in a more pleasant way.

To map the network data to CSML data format (Cell System Markup Language) use the ”csml1.9” or ”csml3.0” buttons according to the required format version.

The ”GIF” button shows the network as image in a new window. This image can be saved then to your machine.

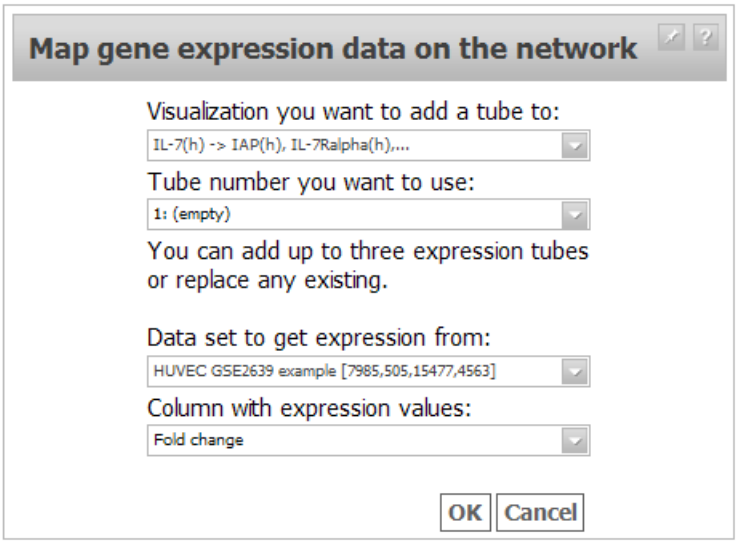
Figure 7.11 Gif image



7.3.1 Adding expression data to the network

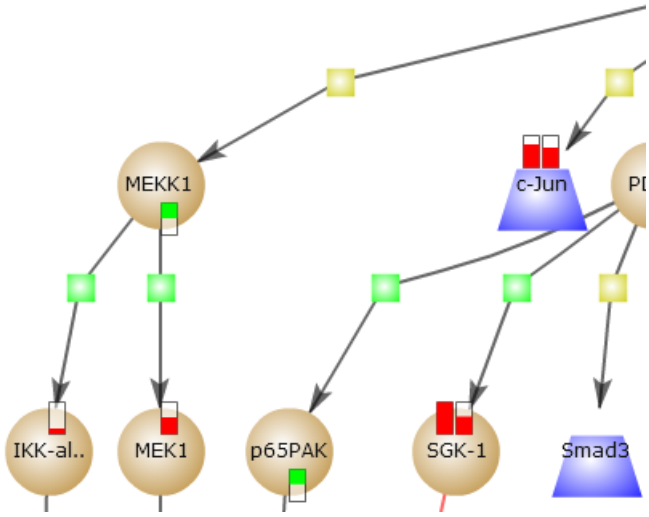
The "Map gene expression" option of the visualized network specific menu provides the ability to add expression information to the network. In a dialog window (Figure 7.12) you can add up to three expression columns from any data set to the network visualized. Select a tube number, a set and an expression column and press the OK button. To add more than one expression column, launch the dialog again or press the button. In the latter case the window will stay open and you will be able to make additional entries. If the tube selected was already assigned to some expression column, it will be reset to the new one.

Figure 7.12 The dialog window


Map gene expression data on the network
Visualization you want to add a tube to:
IL-7(h) -> IAP(h), IL-7Ralpha(h),...
Tube number you want to use:
1: (empty)
You can add up to three expression tubes or replace any existing.
Data set to get expression from:
HUVEC GSE2639 example [7985,505,15477,4563]
Column with expression values:
Fold change
OK
Cancel

The expression will be displayed as a tube (see the figure below), colored in red for positive values and in green for negative ones. The height of the tube represents the numerical value of the expression.

Figure 7.13 Section of a network with expression tubes



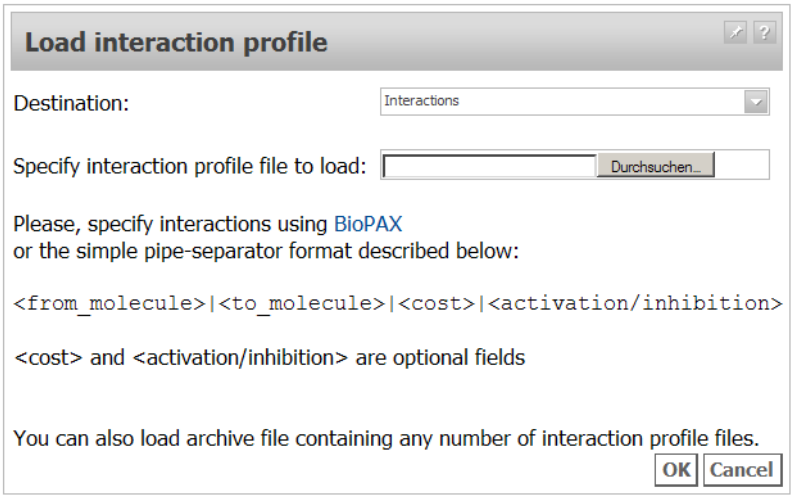
7.4 User defined interactions

7.4.1 Loading of user defined interactions

You have the option to load interactions between molecules into the system and to use these interactions in the search for signaling cascades, key nodes and network clusters.

Select the "Interaction sets" link from the "Load data from file" section of the "File" menu. In the opened dialog window (Figure 7.14) specify the file name and press the OK button. Only molecules that are present in the current BKL release will be loaded.

Figure 7.14 Load interaction profile dialog



Supported file formats

Pipe-separated data files An interaction file should contain the following lines <from_molecule>|<to_molecule>|<cost>|<activation/inhibition>. The direction of the reaction is interpreted as from the first to the second molecule. <cost> and <activation/inhibition> are optional fields. <cost> stands for the cost of the reaction (numerical). <activation/inhibition> can be either 1 or -1 (indicating activation or inhibition respectively).

BioPax data files <**http://www.biopax.org**>. Some examples can be found here <http://pid.nci.nih.gov/PID/browse_pathways.shtml>

Supported databases are listed in Table 3.1.

7.4.2 Interaction representation

Molecule identifiers from BKL and identifiers from the loaded file are displayed as pairs "From - To" with corresponding numerical values, if present.

7.4.3 Joining interaction profiles

Once two or more interaction profiles are uploaded, they can be joined together.

The "Join different interaction sets" operation is accessible from the "Data" menu. Check all interaction profiles you want to join in the dropdown menu and press OK button to invoke the process of creating a joined interaction profile.

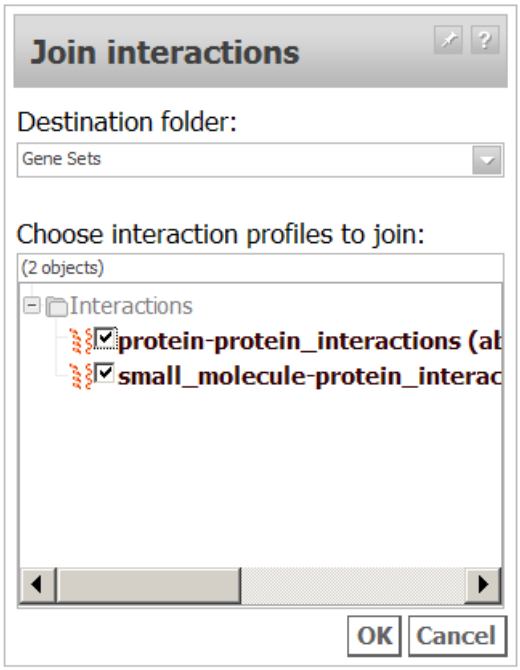
7.5 The BKL database

BKL is a database on gene regulatory and signaling reactions, pathways, and protein attributes, such as expression, phenotype, disease and drug associations. Elements of the relevant signal transduction

Figure 7.15 Interactions dataset

From molecule name	From ID ▲	To molecule name	To ID	Cost	ActRep
SAM50	68653	Noxa	Q13794	1	1
SAM50	68653	PUMA	Q9BXH1	1	1
chic2	74277	AFAP	140935	1	1
chic2	74277	multifunctional protein ADE2	140946	1	1
chic2	74277	RIP	Q13546	1	1
chic2	74277	TNF-alpha	P01375	1	1
chic2	74277	TNFR1	P19438	1	1
chic2	74277	TRADD	Q15628	1	1
chic2	74277	traf2-isoform1	Q12933	1	1
DOK4	114255	ADP	CHEBI:16761	1	1
DOK4	114255	Apaf-1L	O14727	1	1
DOK4	114255	Cytochrome C	P99999	1	1
DOK4	114255	proCaspase-9alpha	P55211	1	1

Figure 7.16 Join interactions dialog



pathways such as hormones, enzymes, complexes, and transcription factors are stored together with information about their interaction.

In the pathways panel of ExPlain, the main components of the BKL database are the molecule, individual reactions and full pathway or reaction chain data. The BKL definitions for these terms are given in the following list.

BKL ENTITIES

Molecule Molecules interact with each other to build pathways. A molecule in BKL is anything that is subject to reactions. Most molecules have a mass, be it a small molecule like ATP, or a protein. No distinction is made between receptors, enzymes, second messengers, transcription factors or other special kinds of proteins. A molecule can also be a group of such entities, like a protein family, a

state of such an entity, like the phosphorylated form or a complex of several other molecules. And finally a molecule can be part of another molecule, either non-covalently bound as in a complex, or covalently bound as in a structural motif of a protein. The reason for such a wide scope for this class is to catch anything that has a specific signaling behavior.

Reaction BKL reactions model interactions, reactions and relationships between molecules. A reaction is a term for all kinds of interactions between signaling entities in signaling or regulatory events. The character of the interaction is more closely defined in its effect field by a set of terms. Reactions, as processes, are not physical entities like molecules, yet they are the central point in the signal transduction database. By representing these reactions between molecules as separate nodes in the graph, it becomes possible to store their properties and annotate them. Since many reactions in signal transduction are catalyzed, and most catalyzed reactions are quasi-unidirectional, all reactions stored in the database are by default unidirectional. Equilibrium reactions are identified in the effect field.

Gene Gene information is linked to the BKL Gene table, where you can find information about the structure of gene regulatory regions, including individual, experimentally demonstrated binding sites for transcription factors, possition weight matrices, ChIP-on-chip and other information.

Genes have been included to provide information about the last step in signal transduction pathways: the regulation of target genes by activated transcription factors. Thus, BKL presents information about complete signaling pathways: starting with the activation of a receptor at the membrane, followed by a cascade of kinases into the nucleus, where a particular transcription factor is activated and regulates the expression of a set of target genes.

Pathway & Chain Pathways reflect canonical pathways for specific signaling molecules (mostly ligands or receptors) and are made up of one or more chains. Chains are sets of consequent reactions, joined by common enzymes or metabolites. Chains can have bifurcations and even loops (if they have a regulatory meaning).

BKL molecules are hierarchically classified into families and can be separately annotated as certain states/modified forms or as components of molecular complexes. Based on their genes and species taxa, molecules are assigned to the groups described below.

TRANSPATH MOLECULE GROUPS

Orthofamily/Family The prefix "ortho" is used when the family entry is not specific for a certain species or higher taxon. An orthofamily is a group entry for a homologous family or superfamily of molecules. A family is a species-specific protein set.

Orthogroup/Isogroup For a single gene, different isoforms such as splice variants may exist. Sometimes in the literature a signaling activity is first attributed to a single molecule, and later it is discovered that there is a whole group of similar molecules. Therefore, a special type of molecule entry is used, which we label "isogroup" for taxon-specific entries and "orthogroup" for orthologous, non-species-specific ones. To these abstracted group entries all the known information can be assigned, when it is not known which specific isoform is involved.

Orthobasic/Basic Molecules of the type "basic" contain data for a specific isoform, e.g. a splice variant, to which an amino acid sequence can be assigned. Again, the prefix "ortho" is used to generalize information for orthologous isoforms from different species.

Orthocomplex/Complex An "orthocomplex" is a group entry for orthologous complexes, consisting of non-covalently bound molecules, where a "complex" describes taxon-specific, non-covalently bound molecules.

The unmodified form of a protein and all its modified forms are its states, where the modification can be by covalent binding, by complex formation, or by change of the environment. The protein per se is a concept that is based on the observation that there is only one gene coding for each protein sequence. All the states share the same gene and consequently part of their structure, the amino acid chain. They are functionally related, often even reversibly transformable into each other. A basic molecule entry captures this concept and is the class of all states of a protein. These states are different molecules, and we store them as different entries. As molecules, they can be used in a pathway assembly. We store general information like the amino acid sequence in the basic molecule entry and link its states to it. In the simplest case there are only two states- an inactive one, and an active one. In other cases, there are more. For example, a transcription factor can be

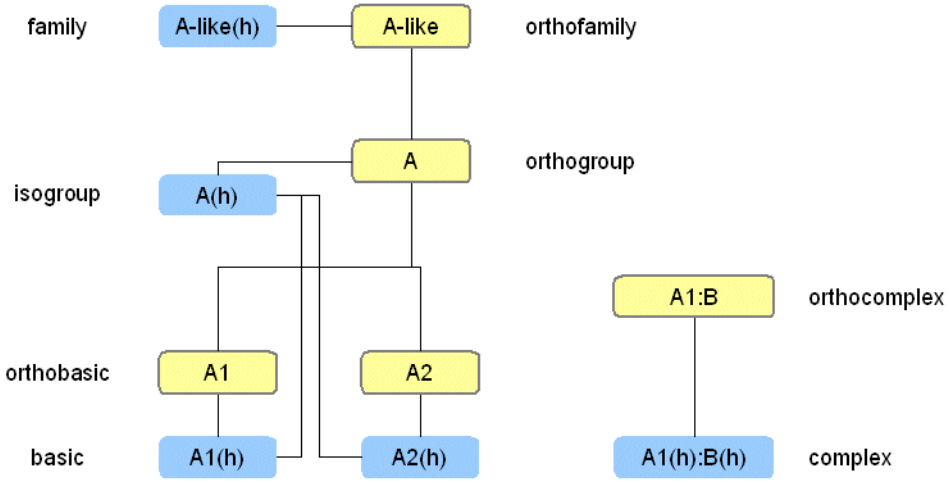
1 de-phosphorylated in the cytosol,

2 phosphorylated in the cytosol, or

3 phosphorylated and bound to DNA in the nucleus,

to name a few possibilities. The same protein will exhibit distinctly different signaling functions in these three states. For example in the first state it will be susceptible to phosphorylation, in the second state to translocation into the nucleus or dephosphorylation and in the third state it will activate transcription. Hierarchical grouping relations and roles of molecules are summarized in Figure 7.17

Figure 7.17 Hierarchical grouping relations between BKL molecules

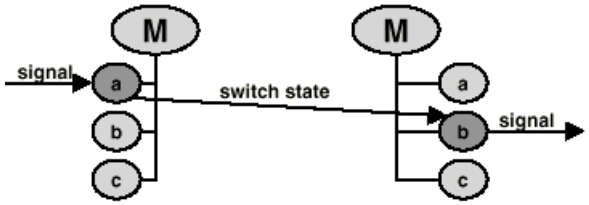


The number of states for a molecule is the product of the number of its modified forms and the number of locations where it is found.

Only compounds which share the same location interact in nature. It is impractical to enter a separate state for each location; Most molecules can be found in several tissues, at several development stages, in several cellular compartments, several organs and several cell types. To enter a state for each possible combination would lead to an explosion in the number of states, and redundancy in the reactions.

This problem is circumvented by using a list of positive and negative locations that is linked to the basic molecule. In each state, the molecule is available only for a subset of all reactions for that molecule. Receiving a signal changes the molecule's state, usually leading to a new state from which reactions are triggered.

Figure 7.18 State switching



Chapter 8

Profiles

This section explains how to create, modify, and import PWM profiles (sets of positional weight matrices of transcription factor binding sites) and the included cut-offs. Cut-offs are threshold values assigned to matrices, and indicate the allowed variation when the matrix is used to predict a binding site. They are measured by comparing the predicted sites with experimentally proven ones. In ExPlain you can choose cut-offs minimizing false positive (overprediction), false negative (underprediction) or the sum of both errors. The created/modified profiles can then be used to search a set of promoter sequences for binding sites as described in Chapter 5, «Transcription factor site search analysis».

8.1 Loading profiles

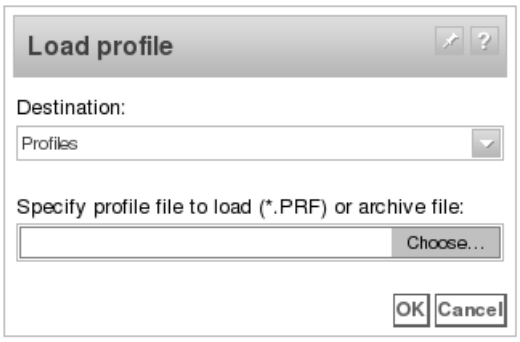
You can add a profile to ExPlain by:

1 Selecting one or several of the user-defined profiles from your TRANSFAC installation, in the data import dialog described in Section 3.1.4.

2 Uploading a local file.

When you choose the "PWM profile" option from the "Load data from file" section of the "File" menu you can load profiles from local source files created from TRANSFAC or from other sources. For a description of how to create profiles in TRANSFAC Professional, e.g. on the basis of a database search result, please consult the documentation for TRANSFAC and MATCH.

Figure 8.1 Load profiles dialog



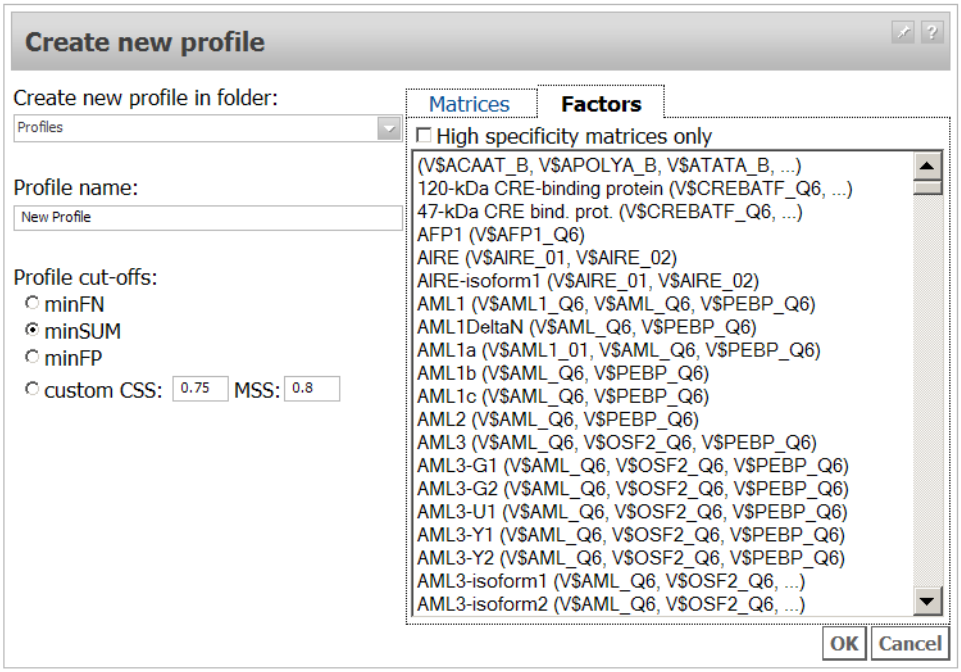
Click on the OK button to start the loading procedure. Imported profiles are added to the "Profiles" folder on the project tree by default. You can select a different folder in the "Destination" field.

8.2 Creating a new profile

8.2.1 New profile dialog

The "Create new profile" dialog, which can be invoked by clicking on "PWM profile" in the "Create new data" section of the "File" menu, provides the functionality to create and modify profiles. Figure 8.2 shows the default mode dialog window, which appears when the "PWM profile" menu link is used and the active tree node is not a profile or gene set. In this mode the dialog window can be used to create a profile from scratch by selecting PWMs from the list box and predefined (minFN, minSUM, or minFP) or custom MSS and CSS cut-offs.

Figure 8.2 The "New profile" dialog window in default mode



"Create new profile in folder: Profiles" is offered by default, but it is possible to choose any other folder. Furthermore you can specify the "Profile name". The contents of the list box and thereby the means of selecting PWMs for the profile can be altered with the "Matrices" and "Factors" tabs above the list box, as well as with the "High specificity matrices only" checkbox.

When the "Factors" tab is selected, transcription factors are shown. Upon profile creation, all PWMs linked to the selected factors are collected in the profile. An entry line contains the factor name, e.g. AhR, and one or more sample identifiers of corresponding PWMs in brackets.

When the "Matrices" tab is selected, the listed entries are TRANSFAC PWM identifiers as shown in Figure 8.3, thus allowing a direct matrix profile compilation.

The "High quality matrices only" option limits the contents of the list to high quality matrices, so that a profile can be created from such PWMs only.

Use the OK button after setting all the necessary parameters to launch the process of profile creation.

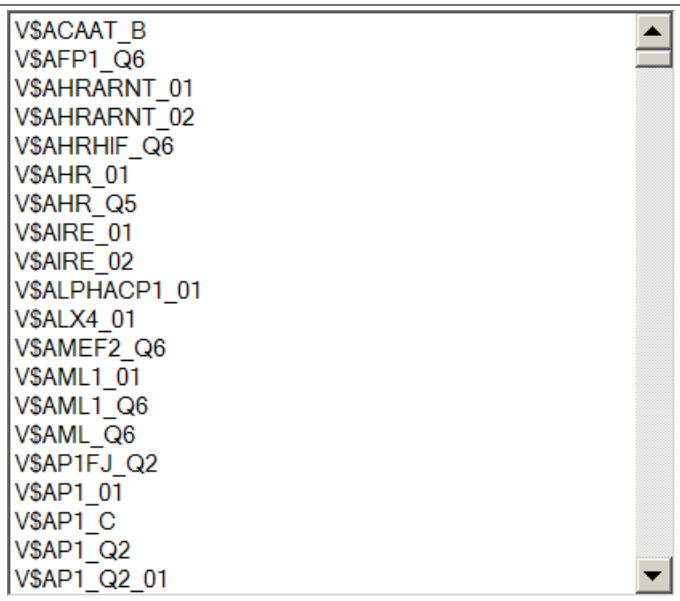
If this dialog window is used in conjunction with an active gene set or profile node, matrix or factor entries of the active profile are preselected in the list box.

8.2.2 Example of profile creation

This example demonstrates a profile creation from a set of up-regulated genes, extracted from the "HUVEC GSE2639 example" geneset in Section 3.3.1. We use the "Create new profile" dialog window to create a profile of all matrices of this gene set.

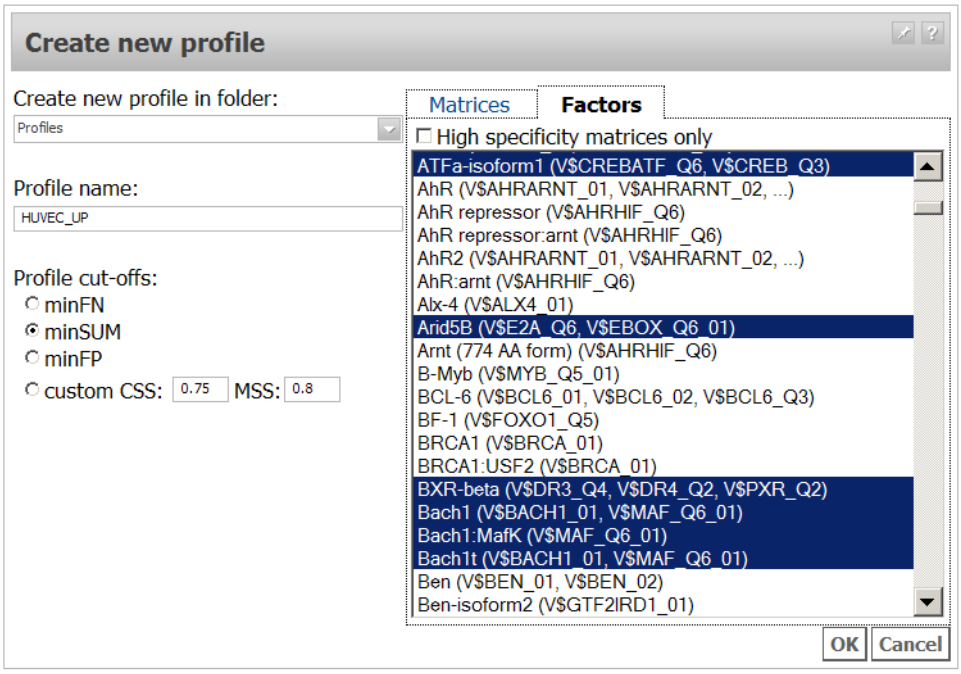
To obtain the profile of matrices from the up-regulated gene set, open the "Create new profile" dialog window, select the rows of the input/output table as shown in Figure 8.4 (if you select nothing, by

Figure 8.3 PWM selector for matrix-based profile compilation



default the system will select all the rows) and press the OK button. It is also possible to invoke the "Create new profile" dialog after selecting the up-regulated gene set. Then all PWMs associated with the gene set are preselected in the list, and no additional matrices need to be selected. Use the "Profile name" field to name the profile "HUVEC_UP" and select the desired destination folder in the field "Create new profile in folder".

Figure 8.4 Input/output table for the creation of the example HUVEC up-regulated profile

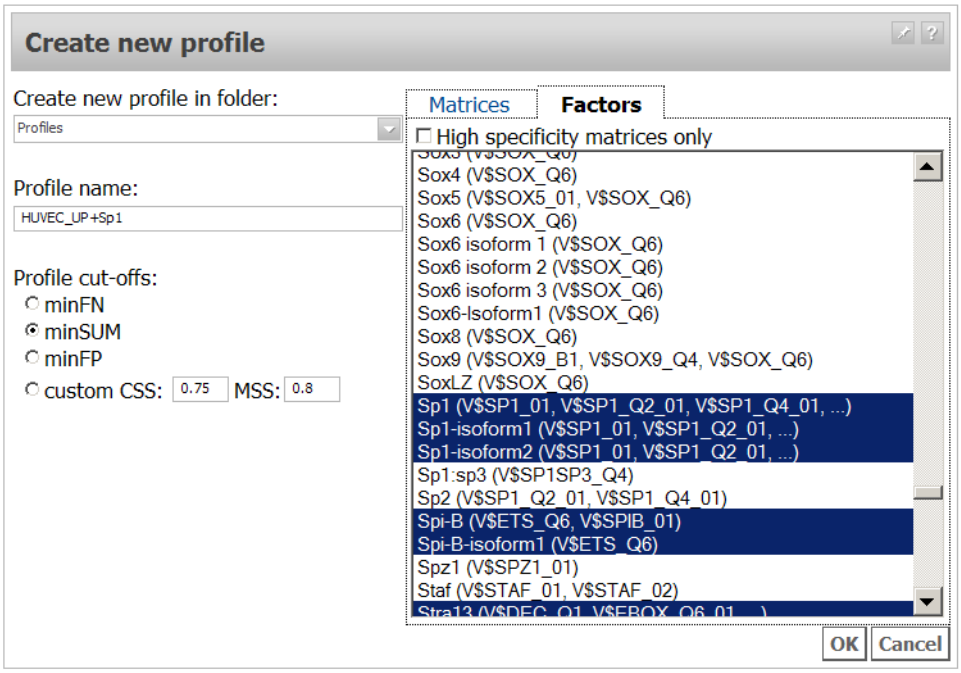


After pressing the OK button, you are redirected to the newly created profile, and the interface changes to the profile mode.

8.2.3 Profile modification

For profile modification choose the profile under study in the tree. Click on the "PWM profile" link in the "Create new data" section of the "File" main menu. Factors/matrices of the chosen profile are then preselected in the list of the dialog window.

Figure 8.5 Dialog window with the active HUVEC up-regulated profile node



In the list box, we extend the "HUVEC_UP" profile by adding matrices of the Sp1 factor. New PWMs are added by scrolling to the corresponding list entries and selecting rows with the mouse pointer while pressing the [Ctrl] key. Since the current "HUVEC_UP" profile was stored with default cut-offs (0.75 as CSS and 0.8 as MSS threshold), we also mark the minSUM radio button, so that matrices of our profile will be configured with the respective cut-offs. Finally, we name the new profile "HUVEC_UP+Sp1". The profile is created by pressing the OK button.

NOTE

The list box supports the use of the left mouse button in conjunction with [Ctrl] or [Shift] as it is standard in many applications. Keep the [Ctrl] key pressed while clicking on items of the list to select several, not necessarily consecutive entries. Alternatively, pressing the [Shift] key marks the range from the previously selected item to the one currently underneath the mouse pointer.

Figure 8.6 shows a table of the profile created by modification of the initial "HUVEC_UP" profile. Each matrix row contains the respective predefined minSUM cut-offs in the CSS and MSS columns. There are additional rows corresponding to Sp1 matrices in the profile table, that were not present in the "HUVEC_UP" profile.

8.3 Creating profiles from gene sets

First select the gene set under study in the project tree. The "Gene set" menu will appear in the menu bar. This menu contains the following options:

1 Minimize false negatives (minFN)

Figure 8.6 The result input/output table of the modified HUVEC up-regulated profile

Matrix name ▲	Recognized factors	Species	CSS	MSS
V$AP1_01	AP-1, FosB, Fra-1, JunB, JunD, c-Fos, c-Jun	Japanese pufferfish, Japanese quail, cattle, chick, dog, domestic pig, gibbon ape, hamster, human, monkey, mouse, rabbit, rat, sheep	0.814	0.88
V$AP1_C	AP-1	chick, clawed frog, fruit fly, gibbon ape, hamster, human, monkey, mouse, rat	0.642	0.841
V$AP1_Q2	AP-1	chick, clawed frog, gibbon ape, hamster, human, monkey, mouse, rat	0.756	0.923
V$AP1_Q2_01	AP-1, FosB, Fra-1, Fra-2, JunB, JunD, c-Fos, c-Jun, deltaFosB	Japanese pufferfish, Japanese quail, cattle, chick, dog, domestic pig, hamster, human, monkey, mouse, rabbit, rat, sheep	0.822	0.951

2 Minimize sums of FP and FN (minSUM)

3 Minimize false positives (minFP)

After clicking on one of these menu links, appropriate for your analysis (for example, "Minimize sums of FP and FN"), a new profile will be created and the interface will change to the profile mode.

Figure 8.7 Input/output table for the profile created from gene set

[+] **Weight matrices profile:** New Profile Rename Cancel

Cut-offs: **minSUM**

Filter (filter bar): none (total 20 rows) Rows per page: 50

Export: Plain text | XLS | RTF | PRF

Mark: All (20) | None | Invert

	Matrix name ▲	Recognized factors	Species	Number of sites	CSS	MSS	»
☐ (*)	V$AHRARNT_01	AhR, AhR2, arnt, arnt-L	guinea pig, human, killifish, mouse, pig, rabbit, rat	24	0.721	0.914	
☐ (*)	V$AHRARNT_02	AhR, AhR2, arnt, arnt-L	guinea pig, human, killifish, mouse, pig, rabbit, rat	25	0.7	0.767	
☐ (*)	V$AHRHIF_Q6	AhR, AhR repressor, AhR repressor:arnt, AhR2, AhR:arnt, Arnt (774 AA form), HIF-1, HIF-1alpha, HIF-1alpha-isoform1, arnt, arnt-L, arnt2	guinea pig, human, killifish, mouse, pig, rabbit, rat	47	0.752	0.979	
☐ (*)	V$ARNT_01	arnt, arnt-L	human, mouse, rat	20	0.712	0.819	

The "Weight matrices profile" field can be used to change the default name "New Profile" to another, appropriate for the specific analysis.

8.4 Profile representation in the result table

The table contains one row for each PWM of the profile. Besides the Matrix name column, each row contains a list of factor names (Recognized factors) column and a Species column. The Number of sites column shows the number of real sequences (suggested binding sites, compiled sequences, TRANSCompel sites) used as the basis for matrix calculation. CSS and MSS columns contain the score thresholds that were stored with the profile.

ExPlain provides you the option to export a profile as a tab-separated text table, Microsoft^{TM} Excel table, Rich Text Format document or Match compatible profile.

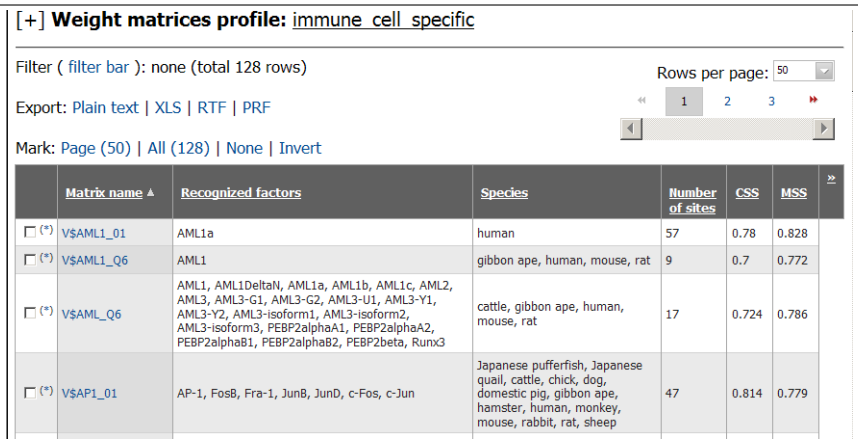
8.5 Profile menu options

If the profile node is active in the project tree, the button "Profile" appears in the main menu. After clicking on this button you will see eight menu links available:

Section: Set Profile - Create profile from selection

Gene set - Create gene set using selected matrices

Figure 8.8 Output table of the "New profile" creation result



[+] **Weight matrices profile:** immune_cell_specific

Filter (filter bar): none (total 128 rows)

Rows per page: 50

Export: Plain text | XLS | RTF | PRF

1 2 3

Mark: Page (50) | All (128) | None | Invert

	Matrix name	Recognized factors	Species	Number of sites	CSS	MSS
(*)	V$AML1_01	AML1a	human	57	0.78	0.828
(*)	V$AML1_Q6	AML1	gibbon ape, human, mouse, rat	9	0.7	0.772
(*)	V$AML_Q6	AML1, AML1DeltaN, AML1a, AML1b, AML1c, AML2, AML3, AML3-G1, AML3-G2, AML3-U1, AML3-Y1, AML3-Y2, AML3-isoform1, AML3-isoform2, AML3-isoform3, PEBP2alphaA1, PEBP2alphaA2, PEBP2alphaB1, PEBP2alphaB2, PEBP2beta, Runx3	cattle, gibbon ape, human, mouse, rat	17	0.724	0.786
(*)	V$AP1_01	AP-1, FosB, Fra-1, JunB, JunD, c-Fos, c-Jun	Japanese pufferfish, Japanese quail, cattle, chick, dog, domestic pig, gibbon ape, hamster, human, monkey, mouse, rabbit, rat, sheep	47	0.814	0.779

Homologous matrices - Extend set of selected matrices by all homologs

Section: Change cut-offs minFN - Minimize false negatives

minFP - Minimize false positives

minSUM - Minimize sums of FP and FN

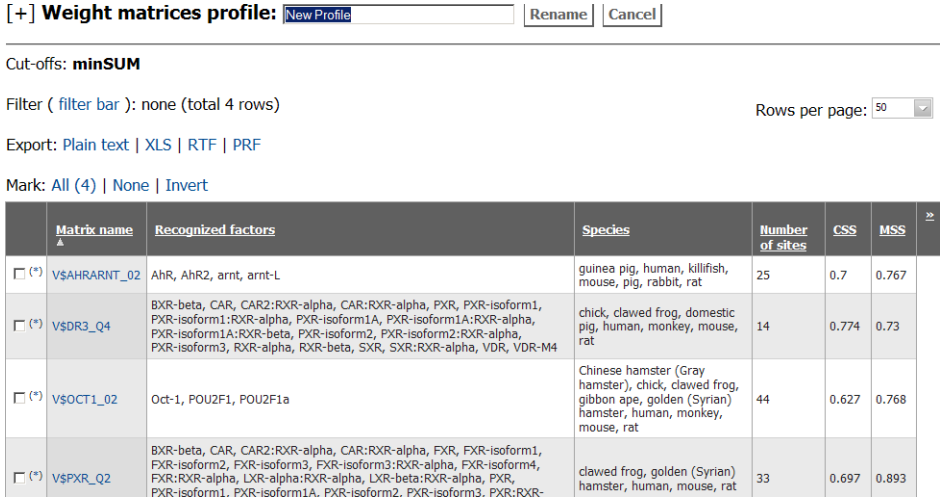
Custom - Set arbitrary cut-offs

Section: Tools Join profiles - Merge several profiles into one

8.5.1 Create profile from selection

When the profile node is active, you can select in the checkbox column all matrices necessary for the analysis. If nothing is selected, all available matrices will be included in the analysis by default. Clicking the menu link "Profile - Create profile from selection" starts the process of profile creation. The new profile is added to the "Profiles" folder of the project tree, and you will be automatically redirected to the profile entry. The name of the newly created profile can be customized in the "User created profile" field editor.

Figure 8.9 Creation of profile from selection output



[+] **Weight matrices profile:** New Profile Rename Cancel

Cut-offs: **minSUM**

Filter (filter bar): none (total 4 rows)

Rows per page: 50

Export: Plain text | XLS | RTF | PRF

Mark: All (4) | None | Invert

	Matrix name	Recognized factors	Species	Number of sites	CSS	MSS
(*)	V$AHRARNT_02	AhR, AhR2, arnt, arnt-L	guinea pig, human, killifish, mouse, pig, rabbit, rat	25	0.7	0.767
(*)	V$DR3_Q4	BXR-beta, CAR, CAR2:RXR-alpha, CAR:RXR-alpha, FXR, FXR-isoform1, FXR-isoform1:RXR-alpha, FXR-isoform1A, FXR-isoform1A:RXR-alpha, FXR-isoform1A:RXR-beta, FXR-isoform2, FXR-isoform2:RXR-alpha, FXR-isoform3, RXR-alpha, RXR-beta, SXR, SXR:RXR-alpha, VDR, VDR-M4	chick, clawed frog, domestic pig, human, monkey, mouse, rat	14	0.774	0.73
(*)	V$OCT1_02	Oct-1, POU2F1, POU2F1a	Chinese hamster (Gray hamster), chick, clawed frog, gibbon ape, golden (Syrian) hamster, human, monkey, mouse, rat	44	0.627	0.768
(*)	V$PXR_Q2	BXR-beta, CAR, CAR2:RXR-alpha, CAR:RXR-alpha, FXR, FXR-isoform1, FXR-isoform2, FXR-isoform3, FXR-isoform3:RXR-alpha, FXR-isoform4, FXR:RXR-alpha, LXR-alpha:RXR-alpha, LXR-beta:RXR-alpha, PXR, PXR-isoform1. PXR-isoform1A. PXR-isoform2. PXR-isoform3. PXR:RXR-	clawed frog, golden (Syrian) hamster, human, mouse, rat	33	0.697	0.893

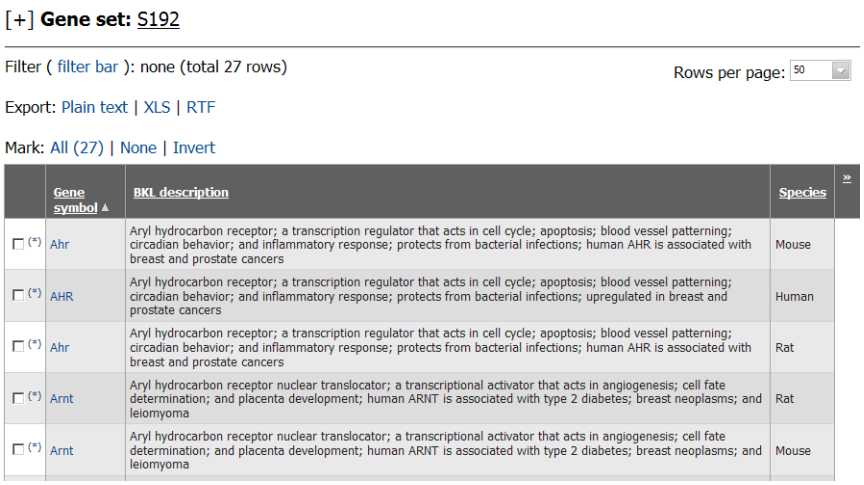
8.5.2 Create gene set using selected matrices

Clicking the menu link "Gene set - Create gene set using selected matrices" starts the process of gene set creation from the matrices selected in the checkbox column (if nothing is selected, all matrices will be included). You will be directed to the newly created gene set, which was added to the project tree under the current profile node.

8.5.3 Extend set of selected matrices by all homologous matrices

Clicking the menu link "Homologous matrices - Extend set of selected matrices by all homologs" starts the process of gene set creation by adding the homologous matrices to the selected ones (if nothing is selected, all profile matrices are included). You will be directed to the newly created gene set, which is added to the project tree under the current profile node.

Figure 8.10 Extention of selected matrices set by homologous matrices output


[+] **Gene set:** S192

Filter (filter bar): none (total 27 rows) Rows per page: 50

Export: Plain text | XLS | RTF

Mark: All (27) | None | Invert

	Gene symbol	BKL description	Species
(*)	Ahr	Aryl hydrocarbon receptor; a transcription regulator that acts in cell cycle; apoptosis; blood vessel patterning; circadian behavior; and inflammatory response; protects from bacterial infections; human AHR is associated with breast and prostate cancers	Mouse
(*)	AHR	Aryl hydrocarbon receptor; a transcription regulator that acts in cell cycle; apoptosis; blood vessel patterning; circadian behavior; and inflammatory response; protects from bacterial infections; upregulated in breast and prostate cancers	Human
(*)	Ahr	Aryl hydrocarbon receptor; a transcription regulator that acts in cell cycle; apoptosis; blood vessel patterning; circadian behavior; and inflammatory response; protects from bacterial infections; human AHR is associated with breast and prostate cancers	Rat
(*)	Arnt	Aryl hydrocarbon receptor nuclear translocator; a transcriptional activator that acts in angiogenesis; cell fate determination; and placenta development; human ARNT is associated with type 2 diabetes; breast neoplasms; and leiomyoma	Rat
(*)	Arnt	Aryl hydrocarbon receptor nuclear translocator; a transcriptional activator that acts in angiogenesis; cell fate determination; and placenta development; human ARNT is associated with type 2 diabetes; breast neoplasms; and leiomyoma	Mouse

8.5.4 Change cut-offs

In the section "Change cut-offs" a user has four options: Minimize false negatives (minFN), Minimize sums of FP and FN (minSUM), Minimize false positives (minFP) and Custom.

Choosing any of the first three menu links will start the process of applying the current choice of score thresholds to the PWMs of the profile without creating a new one. As a result, you will see the same profile with the altered CSS and MSS columns.

After choosing the "Custom" option, a dialog window will appear where you can insert CSS and MSS thresholds according to your choice. Furthermore it is possible to set the cutoffs of the current profile according to the cutoffs of the matrices from another profile by selecting the "use the following profile as template" option and choosing a profile from the corresponding drop down list. You can also set the MSS according to a specific p-value by selecting the "p-value based MSS" option of the "Change profile cut-offs" dialog and chosing an appropriate p-value and species from the respective lists. If the "p-value based MSS" option is used, the CSS will be set to zero. Please also consider that for p-value based cutoffs, matrices for which minimal prediction rate could not be achieved will be removed from the profile, as well as user matrices.

You will be redirected to the profile with the new customized cut-offs.

8.5.5 Merge several profiles into one profile

Clicking on the menu link "Join profiles - Merge several profiles into one" will open a dialog window in which the profiles that will be merged can be selected. After deciding if the highest, lowest, or average cutoffs of the selected profiles shall be used for creation of the joined profile, the process can be started by clicking on the OK button.

Figure 8.11 Custom cut-offs dialog

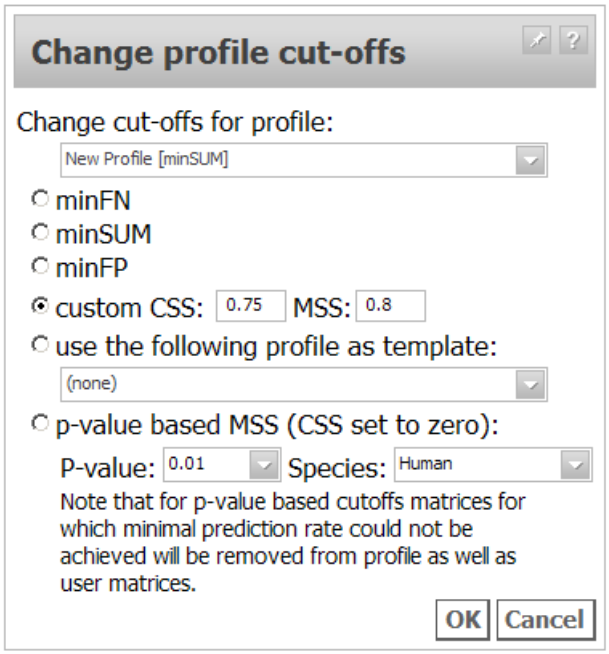
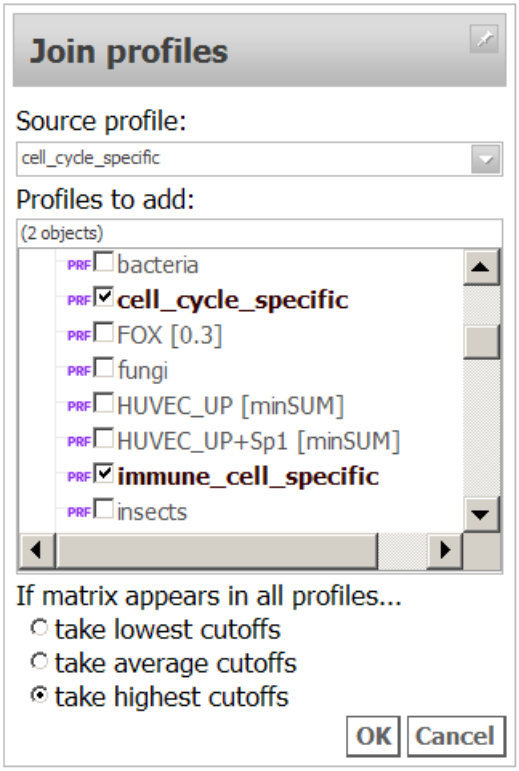

Change profile cut-offs
Change cut-offs for profile:
New Profile [minSUM]
minFN
minSUM
minFP
custom CSS: 0.75 MSS: 0.8
use the following profile as template:
(none)
p-value based MSS (CSS set to zero):
P-value: 0.01 Species: Human
Note that for p-value based cutoffs matrices for which minimal prediction rate could not be achieved will be removed from profile as well as user matrices.
OK
Cancel

Figure 8.12 "Join profiles" dialog



8.6 User matrices

ExPlain provides the ability to create custom positional weight matrices, which will be used to represent transcription factor binding sites. Using a site pattern or sequences alignment, you can create its respective matrix and use it in profiles for site search analysis.

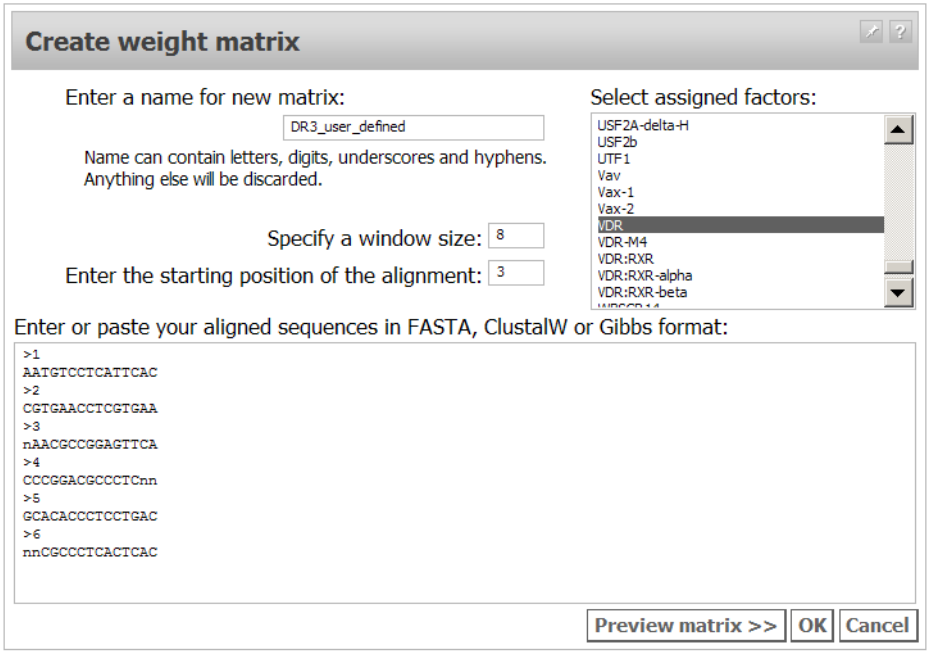
To include a PWM in a profile, first create a matrix as it's described below. Your PWM will then

appear in the matrices list of the profile creation dialog (see Section 8.2) and can be used together with TRANSFAC PWMs.

8.6.1 Creating new matrix

To create a matrix, launch the dialog by clicking on the "Weight matrix" link from the "Create new data" section of the "File" menu. Figure 8.13 shows the dialog window.

Figure 8.13 Weight matrix creation dialog



To create a new matrix you should specify a name using letters, digits, underscores and hyphens. Then you should enter aligned sequences in one of the supported formats. You can change window size and starting position or use default values. You can also assign known transcription factors to the matrix. Select the desired factors in the factor selection control. Note, that you can use [Shift] key button to select range of factors and [Ctrl] key to select several separate factors. As an example we create a matrix named "DR3_user_defined", using a window size of 8 nucleotides. The alignment starts in the third nucleotide, and has the factor VDR assigned.

It is possible to preview PWM and consensus sequence, using the Preview matrix >> button. Figure 8.14 shows the dialog with matrix preview. Using this dialog you can change the matrix name or return to the previous dialog window to change other parameters.

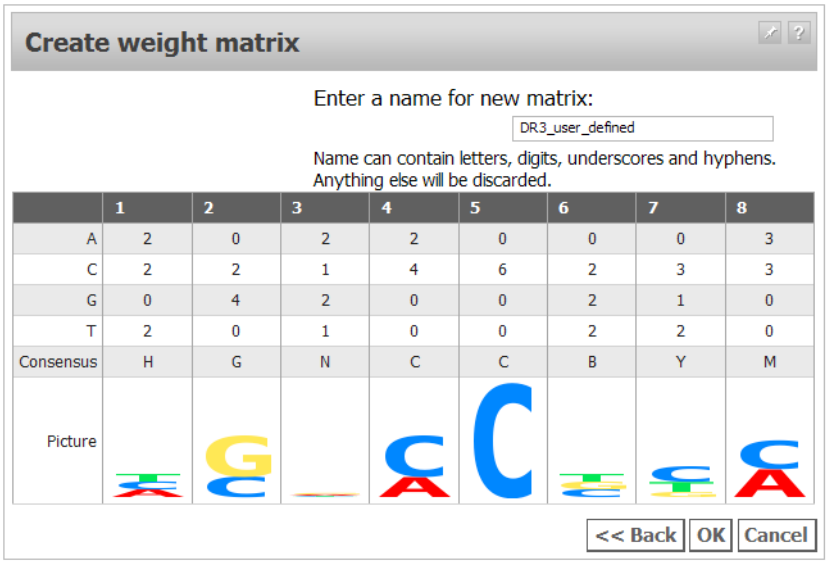
After you set up all required parameters, press the OK button to launch the process of matrix creation. The cut-off values are calculated during the matrix creation process; this can take some time. New matrices are placed in the "Weight matrices" folder on the tree.

8.6.2 Representation of the user matrix

When clicking on the matrix node in the project tree, the matrix will be displayed in the output frame of ExPlain. You can see an example of a newly created matrix in the figure below. The matrix identifier is displayed on the top of the output frame and cannot be changed. The identifier includes an indicator for one of six groups of biological species (V$, vertebrates; I$, insects; P$, plants; F$, fungi; N$ nematodes; B$, bacteria), followed by the matrix name defined by the user in the matrix creation dialog. The PWM is displayed as the nucleotide frequency matrix with one row per nucleotide and one column for each position in the pattern. The derived IUPAC consensus is provided below the frequency matrix. The last row displays the consensus sequence logo.

General matrix information is displayed above the table showing matrix accession ID, matrix quality, window size and transcription factors the matrix is associated with (see next paragraph on factors

Figure 8.14 Weight matrix creation dialog - Matrix Preview



	1	2	3	4	5	6	7	8
A	2	0	2	2	0	0	0	3
C	2	2	1	4	6	2	3	3
G	0	4	2	0	0	2	1	0
T	2	0	1	0	0	2	2	0
Consensus	H	G	N	C	C	B	Y	M
Picture								

changing). An accession ID is given by the ExPlain or BKL matrix generator (for imported matrices). Cut-off values and FP frequency are displayed below the table.

Figure 8.15 Weight matrix

[+] **User matrix:** V$DR3_user_defined

Accession ID: **X18293**
Matrix quality:
Window size: **8**
Binding factors: **VDR** (change)

	1	2	3	4	5	6	7	8
A	2	0	2	2	0	0	0	3
C	2	2	1	4	6	2	3	3
G	0	4	2	0	0	2	1	0
T	2	0	1	0	0	2	2	0
Consensus	H	G	N	C	C	B	Y	M
Picture								

Cut-off values

	CSS	MSS
Cut-offs to minimize false negative matches:	0.844	0.839
Cut-offs to minimize false positive matches:	0.844	1.000
Cut-offs to minimize the sum of both error-rates:	0.844	0.899

FP frequency when tolerating a false negative rate of

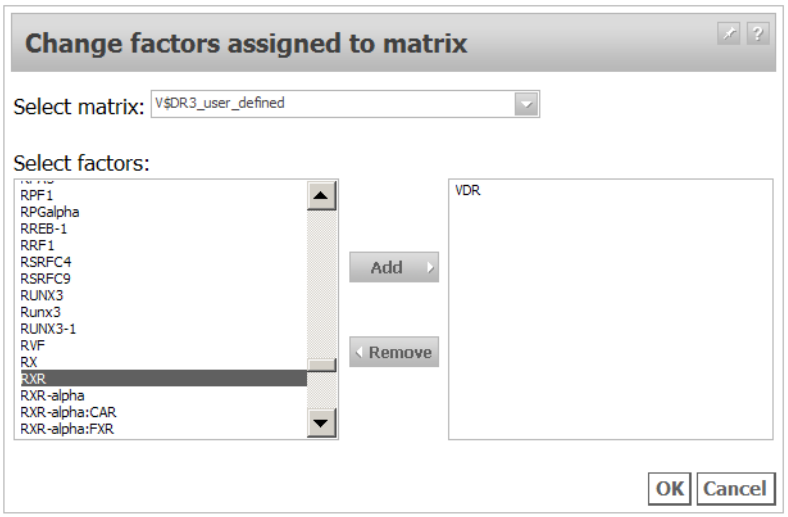
10%	30%	50%	70%	90%
(0.839) 36.878	(0.899) 16.884	(0.923) 9.972	(0.978) 2.895	(0.998) 1.358

8.6.3 Changing factors associated with matrix

To change binding factors associated with a matrix, use the dialog window that can be launched by clicking the ”change” link after the factors list or by using the ”Factors” link in the matrix specific menu. Select a matrix in the drop-down list. Factors already assigned to the matrix are displayed in the left

list, and the right list contains other available factors. Select factors and move them between the two lists using ”Add” and ”Remove” buttons, so that all factors you want to be associated with the selected matrix are collected in the right list.

Figure 8.16 Change factors dialog



8.6.4 Importing user matrices from TRANSFAC

Custom matrices created with the matrix generation tool in TRANSFAC MATCH can also be used in ExPlain application. PWMs from TRANSFAC are imported automatically on the first run of ExPlain and on each run of the new profile dialog. To import matrices manually select the ”Weight matrices” folder in the tree and click » Import matrices from TRANSFAC Match . Imported matrices have the same view as custom created matrices and can be manipulated in the same way.

Chapter 9

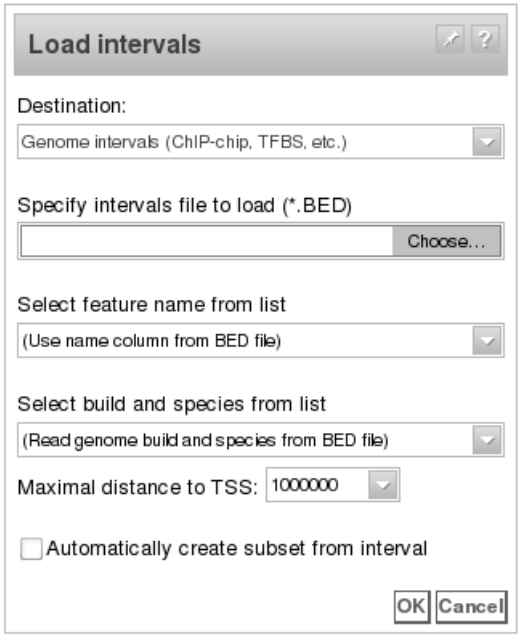
Genome intervals (ChIP-chip, TFBS, ChIP-Seq, Tiling arrays)

9.1 Loading of genome intervals

9.1.1 Loading of genome intervals data from BED file

ExPlain recognizes only the intervals that fall into the range of -10000 .. 1000 nucleotides with respect to the TSS of the gene. To load your interval data, select the ”Load intervals (BED-file)” option from the ”File” menu. Specify a data file using the form shown in Figure 9.1 Besides the file name, you can specify a feature name and a genome build, or select them to be obtained from the BED file. A list of all transcription factors from TRANSFAC database is provided to link them to your data (this information can be further used for filtering the sites search results). Check the option ”automatically create subset from the interval” to create a set of genes covered by the intervals after loading. In the destination field you can choose the name of the folder in the process tree where your data will be saved.

Figure 9.1 Dialog for loading genome intervals from BED file

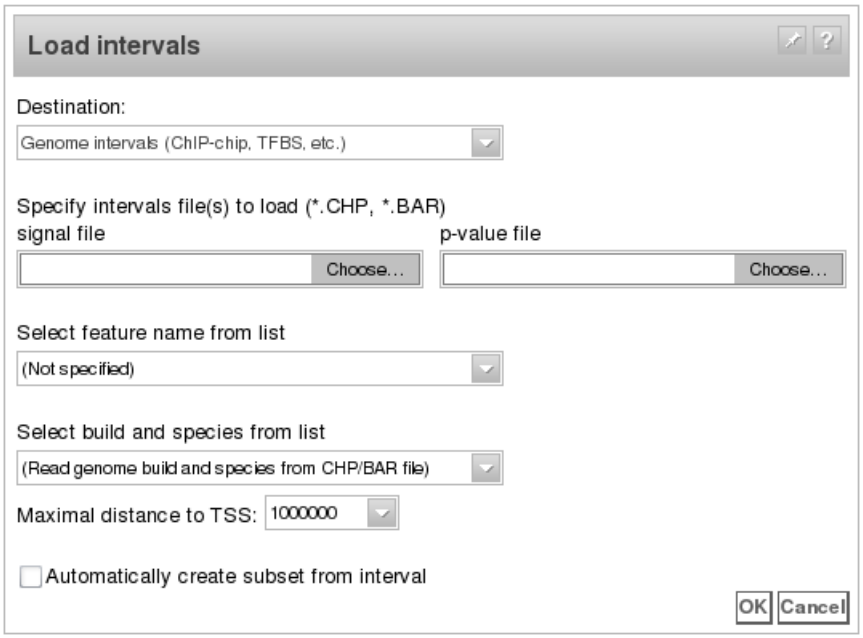


9.1.2 Loading of genome intervals from CHP/BAR file

ExPlain recognizes only the intervals that fall into the range of -10000 .. 1000 nt. from the TSS of the gene. To load your interval data, select the ”Load intervals (CHP/BAR-file)” option from the ”File”

menu. In the dialog window (Figure 9.2) you can specify files with signal and/or p-value columns. You can specify a feature name and a genome build, or select them to be read from the BED file. A list of all transcription factors from TRANSFAC database is provided to link it with your data (this information can be further used for filtering of the sites search results). Check the option "automatically create subset from the interval" to create a set of genes covered by the intervals after loading.

Figure 9.2 Load intervals (CHP/BAR-file) dialog



9.1.3 Loading of genome intervals from Illumina BED file

ExPlain supports import and analysis of Next Generation Sequencing (NGS) data such as output from the Illumina Genome Analyzer in BED format. For peak detection, the novel Model-based Analysis of ChIP-Seq (MACS) algorithm is used. The input in BED format should include the 6th column containing the strand information as required by MACS. The 4th and the 5th columns are not used by MACS, nevertheless some values should be present there. When there are replica BED files from the same experiment, pasting them in one single file will work best for you. ExPlain supports import of archive data in ZIP or GZIP format, which significantly speeds up the process.

To load your NGS data, select the "Load intervals (Illumina BED-file)" option from the "File" menu. In the dialog window (Figure 9.3) you have to specify the genome build and the cutoff distance to the promoter TSS. Only the chip-seq intervals that fall into the range of -10000 .. 1000 from the TSS of the gene will be included in further calculations, however, the neighbor genes, detected at the chosen cutoff, can be exported.

You might want to modify the cutoff p-value used in the peak detection, where more stringent criteria lead to detection of a smaller number of peaks. The MFOLD parameter is used to select the regions with MFOLD fold tag enrichment against a background to build the peak model where the default is 32. The higher the MFOLD, the less of a number of candidate regions will be identified. If you see an ERROR or CRITICAL message from MACS, lowering this parameter is recommended.

9.1.4 Intervals representation

All intervals are grouped by corresponding gene promoters. Species, TSS chromosome location, Gene symbol, promoter ID, and statistical values are given for every line. An example is shown in Figure 9.4 .

Corresponding intervals are accessible through the link in the "entry count" column. By clicking in the entry count number, a new window (Figure 9.5) with all intervals for the current promoter is shown, containing corresponding values and relative gene positions. The picture above the intervals graphically represents the distribution of score and/or p-value if they are present in the initial data.

Figure 9.3 Load intervals (Illumina BED-file) dialog

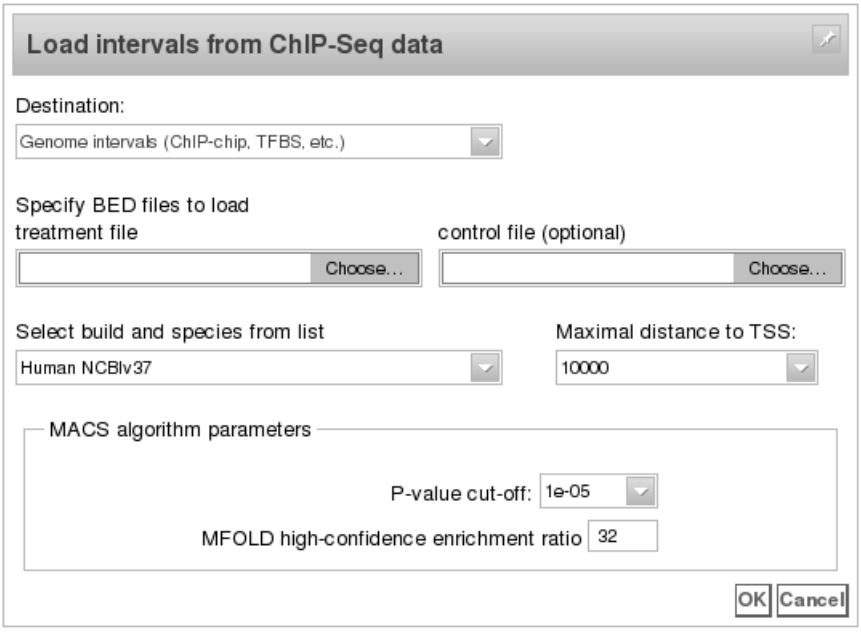


Figure 9.4 Genome intervals dataset, exemple set of known TF binding sites.

[+] Chip-on-Chip data: chr22_chip

Filter (filter bar): none (total 114 rows)

Export: Plain text | XLS | RTF | BED

Mark: Page (100) | All (114) | None | Invert

	Gene symbol	Description	TRANSPro ID	TSS	Species	Entry count
(*)	ASCC2	activating signal cointegrator 1 complex subunit 2	HSA_29339_3	Chr.22 28539402 (−)	Human	2
(*)	ASCC2	activating signal cointegrator 1 complex subunit 2	HSA_29339_4	Chr.22 28536277 (−)	Human	2
(*)	C22orf25	chromosome 22 open reading frame 25	HSA_54792_1	Chr.22 18384296 (+)	Human	2
(*)	COMT	catechol-O-methyltransferase	HSA_1751_1	Chr.22 18309318 (+)	Human	2
(*)	GCAT	glycine C-acetyltransferase (2-amino-3-ketobutyrate coenzyme A ligase)	HSA_4560_2	Chr.22 36543822 (+)	Human	2
(*)	GSTTP1	glutathione S-transferase theta pseudogene 1	HSA_29282	Chr.22 22677258 (−)	Human	2
(*)	GTPBP1	GTP binding protein 1	HSA_9155	Chr.22 37431895 (+)	Human	2

9.2 Recombining intervals

Once the set of intervals is uploaded, it can be reorganized in the following ways: filtered using signal strength, neighbor intervals can be joined, or small intervals can be omitted.

9.2.1 Filtering intervals by conditions

Interval set operations are accessible by selecting an interval set node from the project tree and navigating to the ”Filter interval set” item of the ”Intervals” menu. Figure 9.6 shows the dialog window where you can specify signal conditions and interval parameters. The dialog includes two identical condition fields. The second condition is collapsed by default and is not taken into account. When both conditions are expanded, their parameters can be connected by ”and” or ”or” rules. You can use up to two columns to specify the intervals you wish to extract.

Subsequent lists are used to set a requirement for each marked column. In our example, we seek all

Figure 9.5 Interval entries from a selected gene promoter

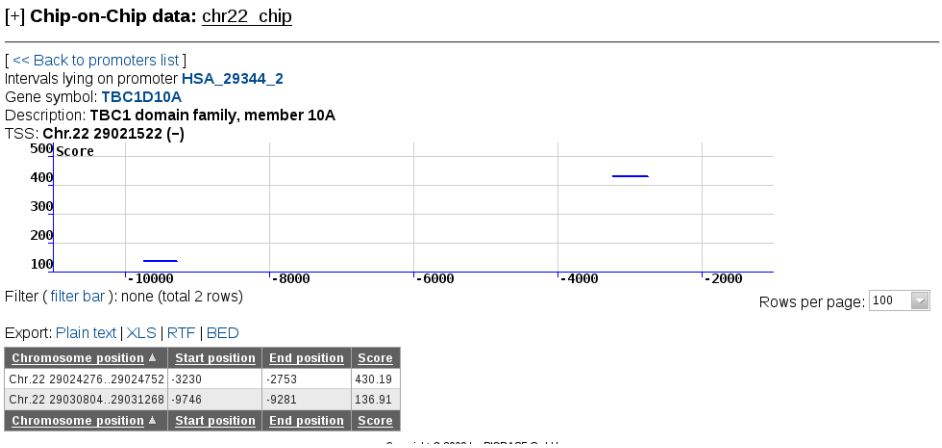
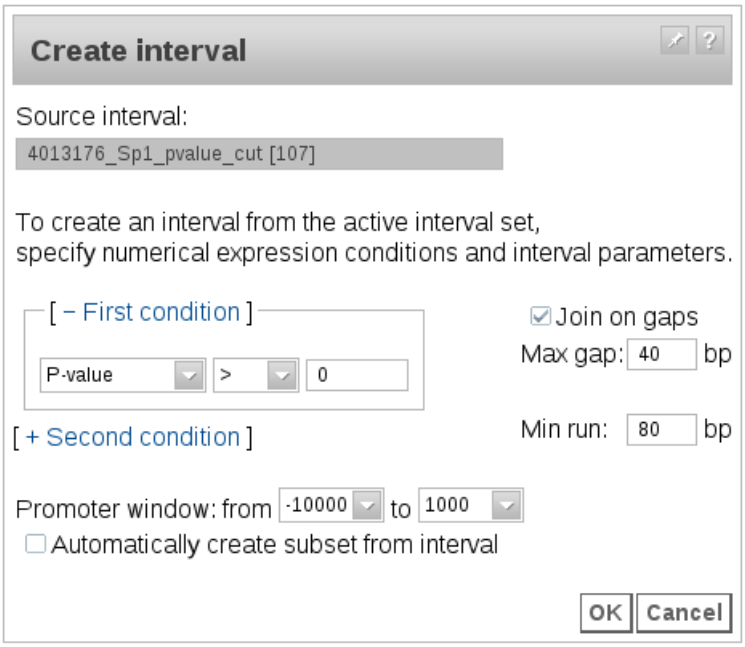


Figure 9.6 Operation on intervals



intervals that have a p-value greater than 0.3. If you have no signal or p-value information attached to your intervals data set, then the above fields will be omitted. The ”Max gap” field accepts the maximum length in bp between intervals. Intervals that are closer will be joined, if the ”Join on gaps” checkbox is checked. The ”Min run” field accepts the minimum length in bp of intervals, and intervals that are shorter will be excluded. To select intervals within a specific position, use the ”filtering by promoter” option. Set up start and end positions from the TSS. Only intervals present in the required window will be shown in the result set.

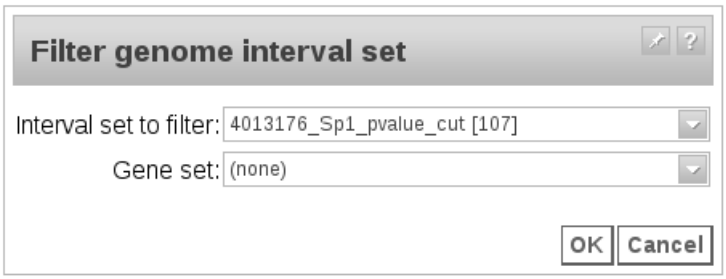
9.2.2 Filtering intervals by gene set

A set of intervals can be filtered by a list of promoters from some gene set. Using the ”Filter interval by gene set” menu option will launch the filtering dialog. Only intervals belonging to promoters from the selected gene set will remain in the result interval set.

9.3 Filtering of MATCH results using intervals

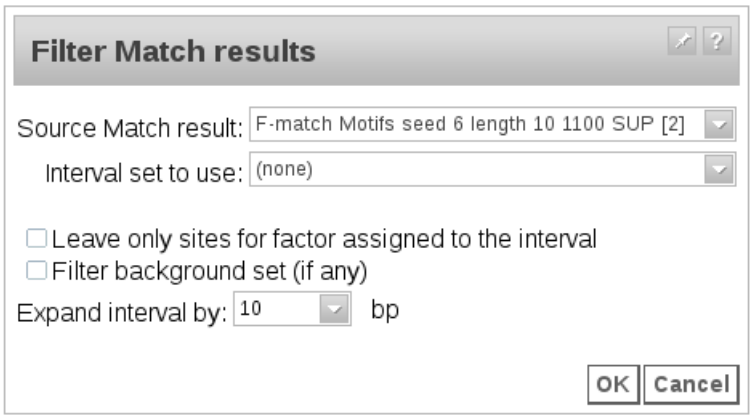
You can use interval sets (Section 9.1) to filter your site search results.

Figure 9.7 Operating on intervals



From the "Analyze" menu, select "Filter site search results using interval set" . If you are standing on a site search result node in the process tree, then it will be chosen in the "Source Match result" field by default, but you can select any other result available from the drop down tree list. The list of previously loaded intervals is available in the "Interval set to use" field, see Figure 9.8. After filtering you will get only binding sites that are found within the intervals selected. If the option "Leave only sites for factor assigned to the interval" option is checked, the results will contain only sites predicted for the factor named at the corresponding interval. To leave sites that don't exactly fit the interval but are close to it, use the "Expand interval" option. Enter here a number of base pairs by which to lenghten the intervals to the left and right. the "Filter background set (if any)" option allows you to control background set filtering.

Figure 9.8 Filter Match results dialog



The OK button launches the process of filtering. It creates a new node that contains the filtered MATCH results.

9.4 Theoretical background of Illumina BED files processing

ChIP-Seq tags represent the ends of fragments in a ChIP-DNA library and are often shifted towards the 3′ direction to better represent the precise protein-DNA interaction site. The size of the shift is, however, often unknown to the experimenter. Since ChIP-DNA fragments are equally likely to be sequenced from both ends, the tag density around a true binding site should show a bimodal enrichment pattern, with Watson strand tags enriched upstream of binding and Crick strand tags enriched downstream. MACS takes advantage of this bimodal pattern to empirically model the shifting size to better locate the precise binding sites. Given a sonication size (bandwidth) and a high-confidence fold-enrichment (MFOLD), MACS slides two bandwidth windows across the genome to find regions with tags more than MFOLD enriched relative to a random tag genome distribution. MACS randomly samples 1,000 of these high-quality peaks, separates their Watson and Crick tags, and aligns them by the midpoint between their Watson and Crick tag centers if the Watson tag center is to the left of the Crick tag center. The distance between the modes of the Watson and Crick peaks in the alignment is defined as 'd', and MACS shifts

all the tags by d/2 toward the 3' ends to the most likely protein-DNA interaction sites.

For experiments with a control, MACS linearly scales the total control tag count to be the same as the total ChIP tag count. Sometimes the same tag can be sequenced repeatedly, more times than expected from a random genome-wide tag distribution. Such tags might arise from biases during ChIP-DNA amplification and sequencing library preparation, and are likely to add noise to the final peak calls. Therefore, MACS removes duplicate tags in excess of what is warranted by the sequencing depth (binomial distribution p-value 10-5). With the current genome coverage of most ChIP-Seq experiments, tag distribution along the genome could be modeled by a Poisson distribution. The advantage of this model is that one parameter, λBG, can capture both the mean and the variance of the distribution. After MACS shifts every tag by d/2, it slides 2d windows across the genome to find candidate peaks with a significant tag enrichment (Poisson distribution p-value based on λBG, default 10-5). Overlapping enriched peaks are merged, and each tag position is extended d bases from its center. The location with the highest fragment pileup, hereafter referred to as the summit, is predicted as the precise binding location. In the control samples, we often observe tag distributions with local fluctuations and biases. Many possible sources for these biases include local chromatin structure, DNA amplification and sequencing bias, and genome copy number variation. Therefore, instead of using a uniform λBG estimated from the whole genome, MACS uses a dynamic parameter, λlocal, defined for each candidate peak as:

λlocal = max(λBG, [λ1k,] λ5k, λ10k)

where λ1k, λ5k and λ10k are λ estimated from the 1 kb, 5 kb or 10 kb window centered at the peak location in the control sample, or the ChIP-Seq sample when a control sample is not available (in which case λ1k is not used). λlocal captures the influence of local biases, and is robust against occasional low tag counts at small local regions. MACS uses λlocal to calculate the p-value of each candidate peak and removes potential false positives due to local biases (that is, peaks significantly under λBG, but not under λlocal). Candidate peaks with p-values below a user-defined threshold p-value (default 10-5) are called and ExPlain reports the quality score of each peak using the formula

-10*log(p-value)

Reference: Zhang et al., Genome Biology 2008 (9), R137.

Chapter 10

Sequences

This chapter explains how to load and analyze nucleotide sequences obtained from any source organism within ExPlain.

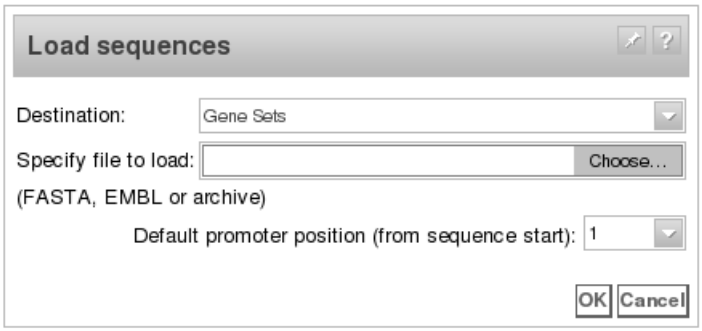
10.1 Loading sequence data into ExPlain

There are two ways to add sequence data to ExPlain. The first si to upload a sequence data file via the ”Load custom sequences” option within the ”File” menu. The second is to directly copy/paste the sequence data into a dialog window via the ”New sequence” option within the ”File” menu.

10.1.1 Load sequence data from a file

To load sequences from a file, use the ”Load custom sequences” menu link within the ”File” menu. A dialog window will open, which allows you to specify the destination folder for the loaded seqeunce(s), the data file to be loaded and the default Transcription Start Site (TSS) position to be used within the sequences. The specified promoter position will be applied to those sequences from the file that do not contain promoter information.

Figure 10.1 ”Load sequences” dialog



By pressing the OK button, sequences from the specified file will be loaded into ExPlain. When a recognized identifier is provided with the sequence (i.e. RefSeq, ENSEMBL, or other identifiers are provided in the header of each FASTA entry, or supported EMBL features are provided, as described below) ExPlain will try to identify the gene associated with each sequence. If no identifying information is found or recognized, ExPlain will create a new user gene entry for each sequence. The sequences will be automatically mapped to a gene set and a new node will appear in the chosen folder on the process tree.

10.1.2 Supported file formats

ExPlain supports three types of sequence files: EMBL and FASTA formatted files and raw nucleotide sequences. You can also upload a compressed archive with sequence files, such as a *.tar or *.gzip file.

For FASTA formatted files, only the identifier and the nucleotide sequence itself will be loaded into ExPlain. This format contains a one line header followed by lines of sequence data. Sequences in fasta formatted files are preceded by a line starting with a ”>” symbol. The first word on this line is the name of the sequence. The rest of the line is a description of the sequence. For example, in the line ”> AATF apoptosis antagonizing transcription factor”, ”AATF” is the name and ”apoptosis antagonizing transcription factor” will be the discription. The remaining lines contain the sequence itself. Blank lines in a FASTA file are ignored, and so are spaces in a sequence. FASTA files containing multiple sequences are just the same, with one sequence listed right after another.

For EMBL formatted files, the following features will be loaded, when provided, into ExPlain: ID (identifier), AC (accessions), DE (description), OS (species), FT promoter or FT exon (TSS position), and SQ (nucleotide sequence).

For raw nucleotide sequences, an identifier will be generated by ExPlain.

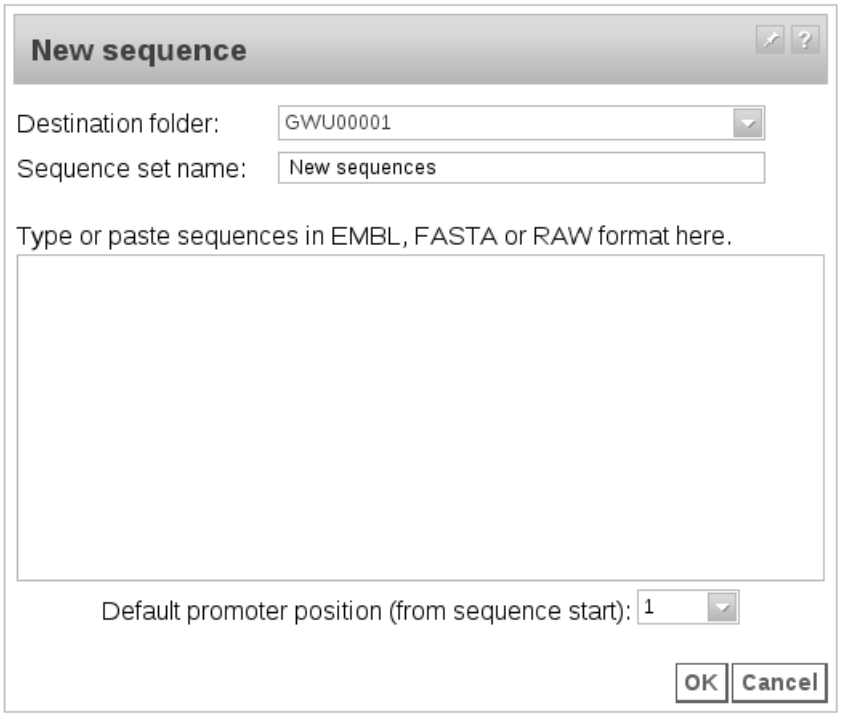
NOTE

Only the first exon start position is taken as the TSS. If both the promoter and the exon features are present in the file, the promoter will be preferred.

10.1.3 Copy/Paste sequence data

To directly copy/paste sequence data into ExPlain use the ”New sequence” menu link within the ”File” menu. A dialog window will open, which allows you to specify the destination folder for the loaded sequence(s), the desired sequence set name and the default TSS position to be used within the sequence(s). Paste sequences in FASTA, EMBL or raw nucleotide sequence format into the text area. By pressing the OK button, the sequences will be loaded into ExPlain in the same manner as described in Section 10.1.1 section above. For more information about supported sequence formats, please see Section 10.1.2 section above.

Figure 10.2 ”New sequence” dialog



10.2 Sequences in ExPlain: example

To understand how loaded sequences are presented in ExPlain, let us consider an example in EMBL format. Consider the file example1.embl which contains the following lines:

Figure 10.3 Example of sequence file (fragment)

```
ID   Y00483; SV 1; linear; genomic DNA; STD; HUM; 1733 BP.
AC   Y00483;
DT   02-APR-1988 (Rel. 15, Created)
DT   14-NOV-2006 (Rel. 89, Last updated, Version 8)
DE   Human gene for gluthathione peroxidase
KW   glutathione peroxidase; GSHPx gene; peroxidase.
OS   Homo sapiens (human)
FH   Key             Location/Qualifiers
FH
FT   source          1..1733
FT                   /organism="Homo sapiens"
FT                   /mol_type="genomic DNA"
FT                   /cell_type="leukocyte"
FT                   /db_xref="taxon:9606"
FT   mRNA            join(154..719,998..1567)
FT                   /product="glutathione peroxidase"
FT   CDS             join(474..719,998..1357)
FT                   /transl_except=(pos:612..614,aa:Cys)
FT                   /product="glutathione peroxidase"
FT                   /db_xref="UniProtKB/Swiss-Prot:P07203"
FT                   /protein_id="CAB37833.1"
FT                   /translation="MCAARLAAAAAQSVYAFSARPLAGGEPVSLGSLRGKVLLIENVAS
FT                   LCGTTVRDYTQMNELQRRLGPRGLVVLGFPCNQFGHQENAKNEEILNSLKYVRPGGGFE
FT                   PNFMLFEKCEVNGAGAHPLFAFLREALPAPSDDATALMTDPKLITWSPVCRNDVAWNFE
FT                   KFLVGPDGVPLRRYSRRFQTIDIEPDIEALLSQGPSCA"
FT   exon            154..719
FT                   /number=1
FT   intron          720..997
FT                   /number=1
FT   exon            998..1567
FT                   /number=2
XX
SQ   Sequence 1733 BP; 277 A; 529 C; 526 G; 401 T; 0 other;
     aacctagatc cctctgctgt cccctgcact gccggtaaca tggcacagca gagcagggtt        60
     gtttgtgcac gggcagctcc tgcagctgct gccgtcgccc accagcctcc tatgccaaac       120
     cccacatcct aactcaggaa cctctgagaa aaaacggagc cctcgagggg cccagccttg       180
     gaagggtaac tggaccgctg ccgcctggtt gcctgggcca gaccagacat gcctgctgct       240
     ccttccggct taggaggagc acgcgtcccg ctcgcgcgca ctctccagcc ttttcctggc       300
     tgaggagggg ccgagcctcc ggtagggcgg gggccggatg aggcgggacc tcaggcccgg       360
     aaaactgcct gtgccacgtg acccgccgcc ggccagttaa aaggaggcgc ctgctggcct       420
     ccccttacag tgcttgttcg gggcgctccg ctggcttctt ggacaattgc gccatgtgtg       480
     ctgctcggct agcggcggcg gcggcccagt cggtgtatgc cttctcggcg cgcccgttgg       540
```

When this file is loaded to ExPlain, a new sequence node called example1 will appear in the project tree.

Figure 10.4 New item in project tree

Each nucleotide sequence is considered as the sequence of one gene. When supportin ginformation is provided the identifier is taken from the ID line (Y00483 in this case) and is used as the gene symbol in posterior analyses. In the figure below you can see the sequence entry loaded from the example file above. Once loaded, EMBL formatted sequence can be retrieved using the export link.

This sequence (or sequence set, if you upload more than one) is now available in ExPlain for further use.

Figure 10.5 Sequence item

Sequence: Y00483

Delete

[<< Back]
Export: EMBL

Description: Human gene for gluthathione peroxidase
Gene symbol: Y00483
Sequence length: 1733
Spesies: 86
Promoters:

Promoter ID	TSS
USR_104939	1

Nucleotide sequence:

```
aacctagatc cctctgctgt cccctgcact gccggtaaca tggcacagca gagcagggtt        60
gtttgtgcac gggcagctcc tgcagctgct gccgtcgccc accagcctcc tatgccaaac       120
cccacatcct aactcaggaa cctctgagaa aaaacggagc cctcgagggg cccagccttg       180
gaagggtaac tggaccgctg ccgcctggtt gcctgggcca gaccagacat gcctgctgct       240
ccttccggct taggaggagc acgcgtcccg ctcgcgcgca ctctccagcc ttttcctggc       300
tgaggagggg ccgagcctcc ggtagggcgg gggccggatg aggcgggacc tcaggcccgg       360
aaaactgcct gtgccacgtg acccgccgcc ggccagttaa aaggaggcgc ctgctggcct       420
ccccttacag tgcttgttcg gggcgctccg ctggcttctt ggacaattgc gccatgtgtg       480
ctgctcggct agcggcggcg gcggcccagt cggtgtatgc cttctcggcg cgcccgttgg       540
```

Chapter 11

Statistical Analysis of Microarray Data

This chapter describes the statistical tools provided within ExPlain for analysis of Microarray data from Affymetrix® CEL files. The following four steps are described in detail below:

Loading of CEL files / Assignment of factors and levels
Low level analysis
Quality Control
High level analysis
Microarray data analysis.

IMPORTANT NOTICE:

WARNING: This chapter addresses highly specialized statistical tools, if you have doubts about their usage, we recommend you to leave the default values provided by ExPlain when using them.

The statistical tools are available only if the R language software suite is installed on your system.

This ExPlain program allows the analysis of Microarray data from Affymetrix™ CEL files. The workflow includes four steps described in details below. The following order of the steps is required:

Loading of CEL files / Assignment of factors and levels

Low level analysis

Quality Control

High level analysis

11.1 Loading of CEL files and assignment of factor-level information

The CEL files must first be archived in ZIP, TAR, or Gzip TAR format. Correctly archived files can be loaded into ExPlain via the ”File” -> ”Load CEL files” menu option. A dialog window will open which allows you to specify the destination folder and the actual file to be uploaded. Press the OK button to launch the process.

The CEL files are extracted from the ZIP or TAR archive, and the factor-level assignment options are shown (Figure 11.2). You must assign factors and levels for the data by clicking on the [Change configuration] link provided. A dialog window will open in which the factors and levels for the data sets can be specified. (Figure 11.3).

Figure 11.1 "Load CEL files" dialog

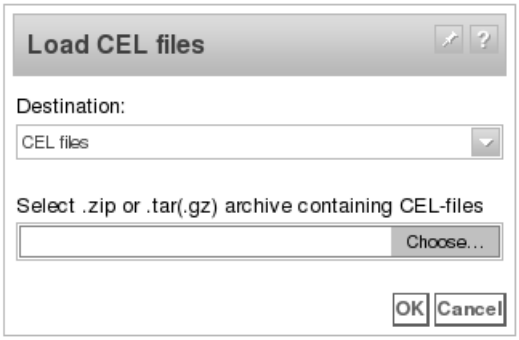
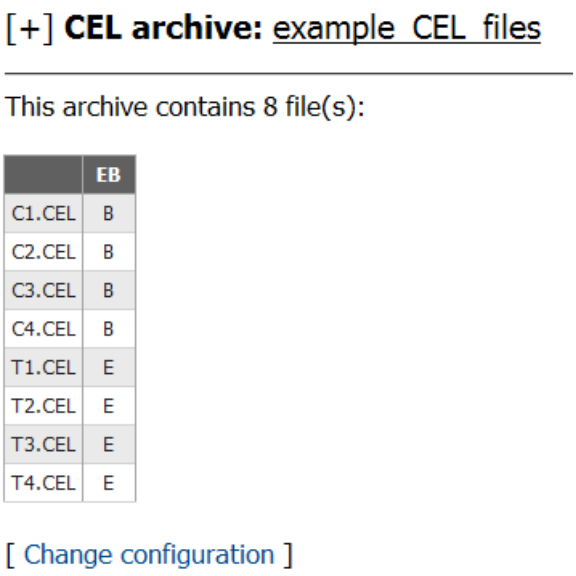


Figure 11.2 Loaded CEL files with assigned factors / levels



11.1.1 Factor level assignment

The factor level assignment dialog (Figure 11.3) is designed for experiments having multiple factors and levels. Factors are added by clicking the [Add factor] button and entering the factor name in the popup window that appears. In order to enter levels for the factor, the factor must be selected in the "Factor:" list box and then the [Add levels] button must be clicked. When a new factor is created, popup for levels will appear automatically.

Assigning CEL files to a particular factor-level combination involves selecting the factor and marking the cells in the assignment table where CEL files are displayed as rows and levels dispalyed as columns. For your convenience, when the cursor is over the cell, the corresponding file-level pair is highlighted. Note, that only one level of a factor can be assigned to a file and this is controlled automatically. Additional factors and their levels can be added in a similar manner. To view the summary table of all assignments, click the "Summary" link. After all factor level assignments have been made, the OK button should be clicked to store the factor-level information in the database.

The factor-level information assigned will be displayed as a table (see figure below). Clicking the [Change configuration] button will launch the factor-level assignment dialog.

11.2 Low level analysis

You have the option to choose between pre-selected techniques for each method, such as MAS4.0, dChip, RMA, GCRMA, or MAS5.0, or to configure advanced options (see below). The following steps will be

Figure 11.3 Factor level assignment interface

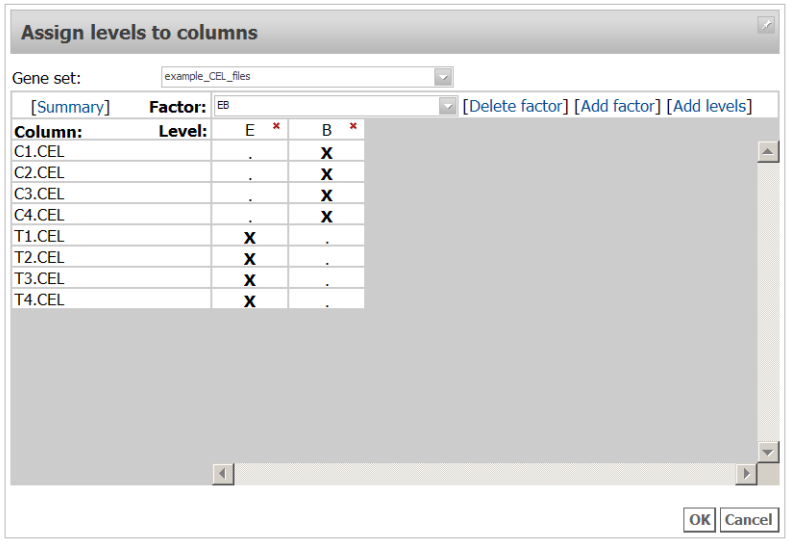


Figure 11.4 Saved factor-level information

	EB	Other factor
C1.CEL	B	high
C2.CEL	B	medium
C3.CEL	B	low
C4.CEL	B	high
T1.CEL	E	low
T2.CEL	E	low
T3.CEL	E	medium
T4.CEL	E	high

[Change configuration]

performed during the low level analysis: background correction, normalization, PM correction, and summarization.

Table 11.1 shows the individual techniques used for the pre-selectable methods MAS 4.0, dChip, MAS 5.0, and RMA.

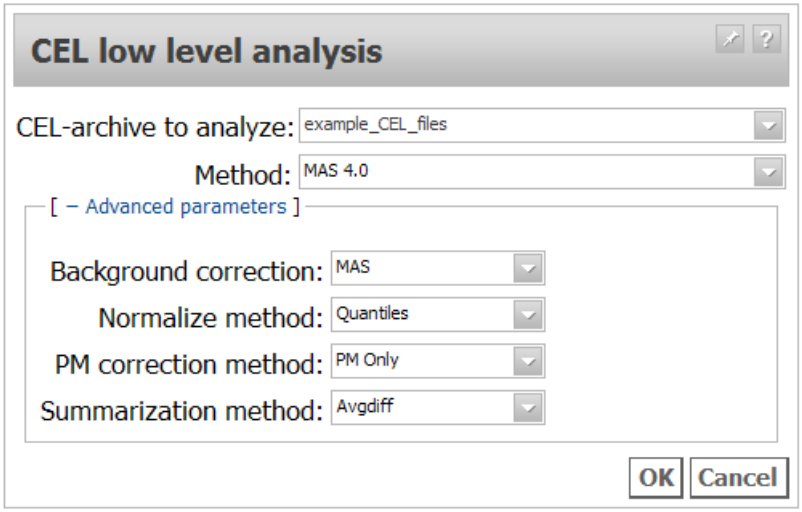
Table 11.1

Pre-selected method	Background correction	Normalize method	PM correction method	Summarization me
MAS 4.0	MAS	Quantiles	PM only	Avgdiff
dChip	MAS	Invariantset	PM only	Liwong
MAS 5.0	MAS	Quantiles	PM only	MAS
RMA	RMA	Quantiles	PM only	Median Polish

Table 11.2 lists the techniques available in ExPlain for each low level analysis step. You can choose any combination of these techniques when performing low level analysis.

The output of a low level analysis is a table containing the microarrays in the columns and the probe IDs in the rows. It may be noted that, from this step onwards, expression value tables generated in each step can be exported as an ExPlain gene set using the "Gene set - Convert selected rows to gene set" menu option.

Figure 11.5 Low level analysis dialog



11.3 Quality control

This step enables you to exclude certain microarrays from further analysis based on quality control tests. BioConductor includes several methods to assist the quality control:

Background statistics This procedure checks whether the background values across the arrays are comparable.

Global scaling factors This standard normalization procedure involves globally rescaling the arrays to set the median probe intensity to the same level. The scaling factors should not differ by more than 3-fold.

Percent present calls Extremely low (below about 30%) or high (above about 60%) values for the percentage of probes have potential quality problems.

3′/5′ ratios Affymetrix includes probes at the 3′ and 5′ ends of some control genes; the ratios should be less than 3.

The quality control tests are launched from the output of low level analysis through the menu option ”CEL” -> ”Quality Control” . To run the quality control tests, you must specify a data set, rules and visuals, and then to click on the OK button (Figure 11.6). With Visuals you will be able to see an overall intensity, distribution of the feature intensities, and RNA degradation for your data. At hte end of the quality control test it is possible to generate a quality control report in PDF format, which shows graphical representations of several quality measures.

The output of the quality control tests is a table containing all CEL files as rows, and the results of each quality control step as columns. To remove failed CEL files from the further analysis, deselect the check boxes of the respective CEL files and click on the ”Submit” button (Figure 11.7). The result of the quality control filters is a table similar to the Low level analysis output, but without the deselected microarrays.

Clicking the View button in the Quality Report section below the result table opens a report in PDF format that displays several quality control features. Alternatively, the report can be is generated with the affyQCReport package from Craig Parman and Conrad Halling [[29]] and can also downloaded by clicking on the Download button. The following description of the content of this report is based on the affyQCReport manual provided by the package authors.

Figure 11.6 Quality control dialog window

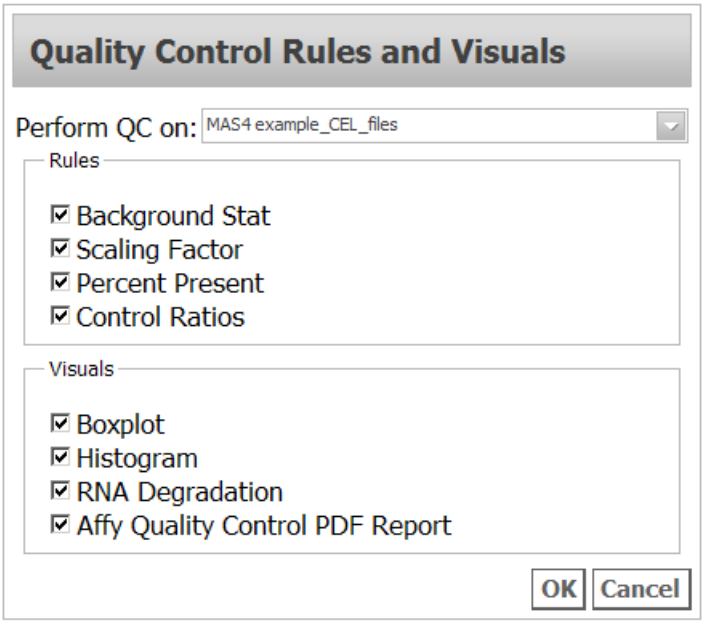


Figure 11.7 Quality control output

		Background Stat	Scaling Factor	Percent Present	Control Ratios
e1.CEL	☑	FAIL	PASSED	PASSED	PASSED
e2.CEL	☑	FAIL	PASSED	PASSED	PASSED
e3.CEL	☑	FAIL	PASSED	PASSED	PASSED
e4.CEL	☑	FAIL	PASSED	PASSED	FAIL
e5.CEL	☑	FAIL	PASSED	PASSED	FAIL
e6.CEL	☑	FAIL	FAIL	FAIL	FAIL

Quality Report (PDF): Show or Save

Plot(s):

[+ Boxplot]

[+ Histogram]

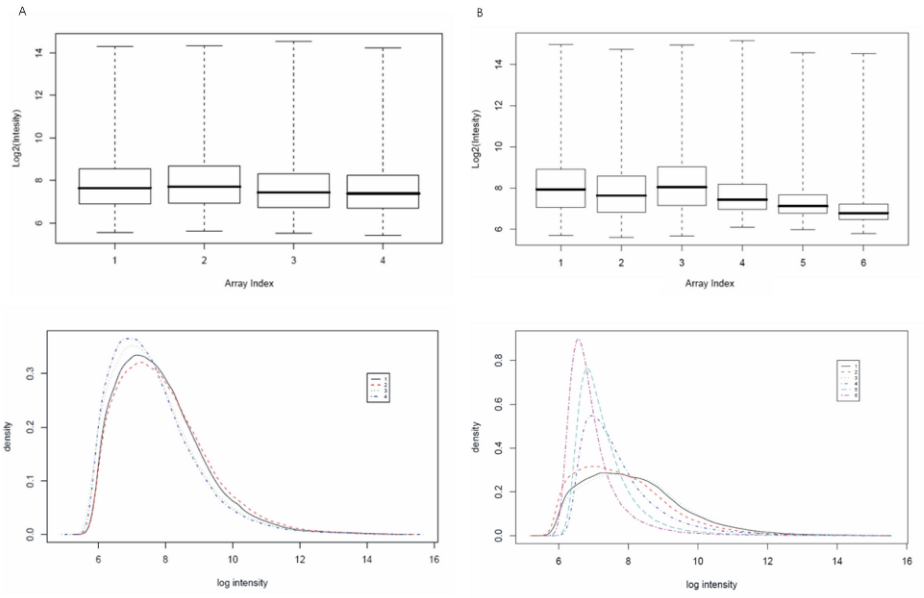
[+ RNA Degradation]

Submit

Page 1: Overview of all analyzed arrays. A table giving the internal numbering used in the plots of the quality control report and the corresponding array names is displayed.

Page 2: The plot at the top of the page is a boxplot of the all pm intensities, which enables an analysis of the overall probe intensities of the arrays (see also ”Boxplot” section below). Arrays with a low average intensity need to be thoroughly examined. The second plot shows the kernel density estimates of the all pm intensities. Special attention should be paid to the shape of the density curves. Arrays, which show a significantly different density profile are suspect.

Figure 11.8 Plots of page 2 of the affyQCReport for datasets that are of reasonable quality (A) and bad quality (B)



Page 3: Plot of the 3′ : 5′ ratios, percent present calls, and average background levels. The 3′: 5′ ratios are derived for spiked-in and control genes specific to the array type. All data within the plot that do not fulfil the recommended criteria are marked in red, otherwise they are colored blue. A detailed description of the plot can be found in the document "QC and Affymetrix data" [[30]] from the documentation of the simpleaffy package [[31]]. Important parts of the description provided in this document are given as excerpts below. The figure is plotted from the bottom up with the first chip being at the base of the diagram. Dotted horizontal lines separate the plot into rows, one for each chip. Dotted vertical lines provide a scale from -3 to 3. Each row shows the percent present, average background, scale factors and GAPDH / beta-actin ratios for an individual chip.

GAPDH 3′ : 5′ values are plotted as circles. According to Affymetrix they should be about 1. GAPDH values that are considered potential outlier (ratio $>$ 1.25) are coloured red, otherwise they are blue.

Beta-actin, 3′ : 5′ ratios are plotted as triangles. Because this is a longer gene, the recommendation is for the 3′ : 5′ ratios to be below 3; values below 3 are coloured blue, those above red.

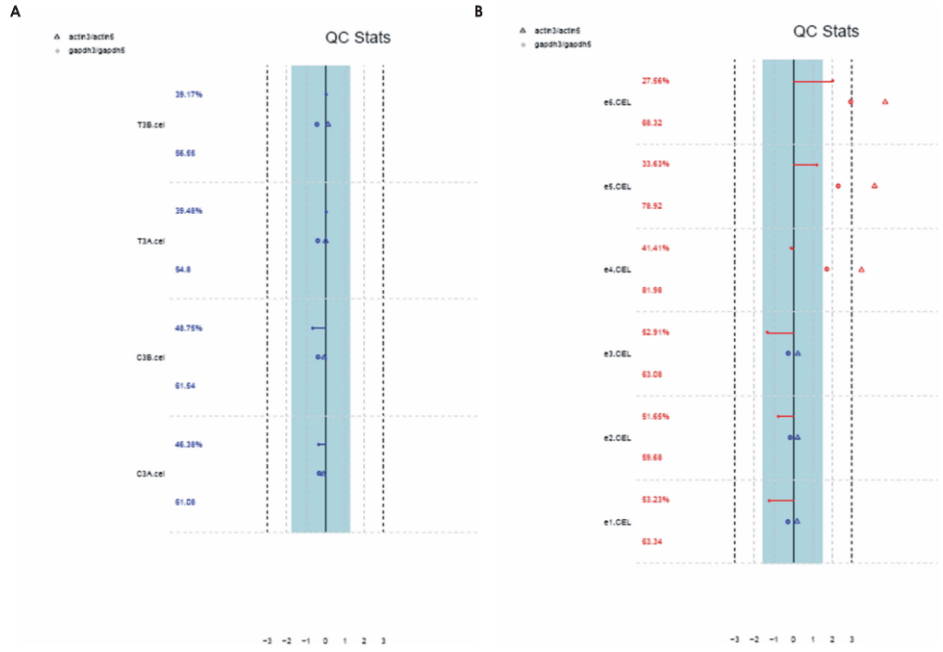
The blue stripe in the image represents the range where scale factors are within 3-fold of the mean for all chips. Scale factors are plotted as a line from the centre line of the image. A line to the left corresponds to a down-scaling, to the right, to an up-scaling. If any scale factors fall outside this '3-fold region', they are all coloured red, otherwise they are blue.

Percent present and average background, are listed to left of the figure.

Page 4: Box plots of the intensities of the positive and negative control elements from the affymetrix arrays. The average values and standard deviations of the intensities should be similar for all arrays of the data set.

Page 5: Center of intensity (COI) plots for positive and negative control elements. Ideally all COIs will be located in the center of the plot (0,0). Some variation to the COI can however be expected and tolerated. If the COI of an array has a coordinate equal to or larger than 0.5 the corresponding dot is labelled with the array identifier. A significant shift of the COI far from the center indicates a spatial variation in the hybridization.

Figure 11.9 Page 3 of the affyQCReport: Plot of the 3′ : 5′ ratios, percent present calls, and average background levels for data sets that are of reasonable quality (A) and bad quality (B).



By clicking on the plus buttons in the Plot(s) section below the quality control result table the user can display the graphs of different quality control features.

Boxplot: This graph enables an analysis of the overall probe intensities of an individual array and a comparison of the intensities between the different arrays of the user's data set. The upper and lower borders of each box indicate the 75th and 25th percentiles in the distribution of the intensities, and the black bar within a box marks the median. The lines extending from each box show the spread of the intensity values. There should be no large differences in the levels of the raw probe intensities. Arrays with a significantly different range of intensity values or a low average intensity are suspect and need a careful examination.

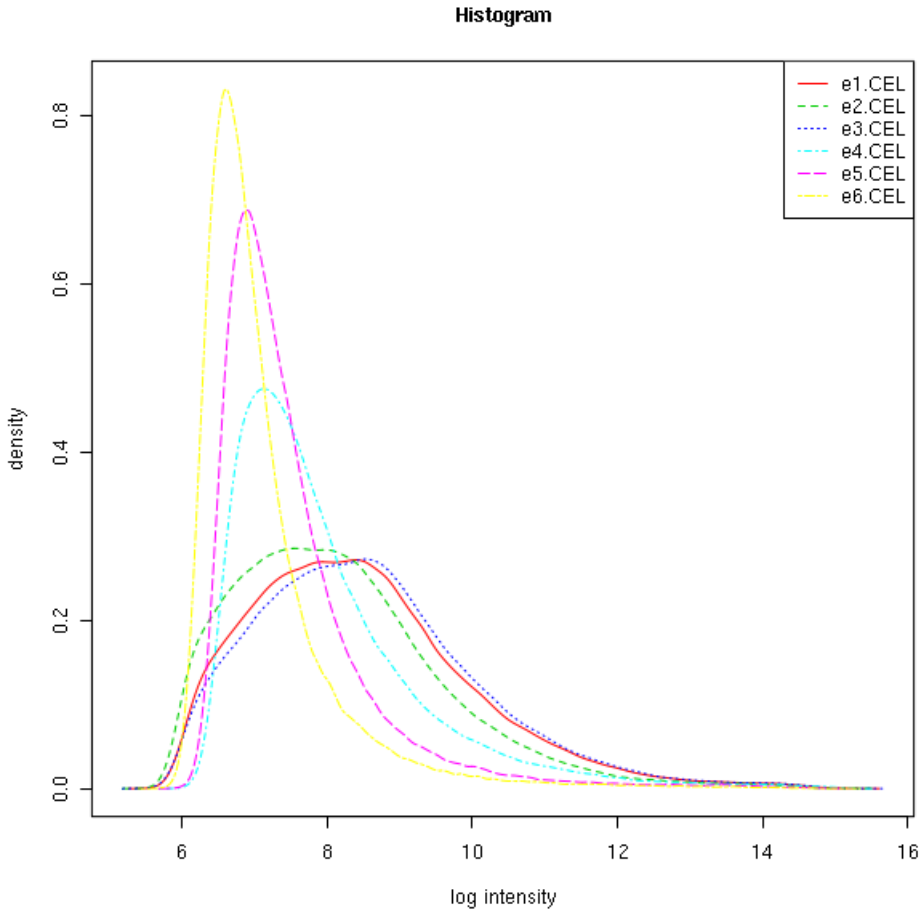
Histogram: Kernel density plot of the probe intensities. The intensities of the arrays are represented by individual lines. The legend in the upper right corner shows the line to array assignment. Differences in the shape or center of the distribution indicate a need for thorough normalization.

RNA Degradation: The plot shows the (shifted and scaled) mean intensity from the 5′ to the 3′ end of the mRNA. Each line in the graph corresponds to one microarray of the analysed data set. Since the RNA is degraded from the 5′ to the 3′ end, intensities at the 5′ end are lower than those at the 3′ end. The slopes and shapes of the lines should be similar for all arrays. DNA-chips with significantly different profiles and slopes should be carefully examined. High slopes of the plotted lines indicate a poor quality due to RNA degradation or other factors that cause systematically elevated intensities at the 3′ compared to the 5′ end.

11.4 High level analysis

High level analysis allows you to find out differentially expressed genes. There are four high level analysis techniques for studying single data sets available for use in ExPlain: ANOVA, Fold change, Empirical Bayes and Generalized Linear Model.

Figure 11.10 Kernel density plot of the probe intensities for a data set, that is of bad quality.



Although fold change is presented as a separate technique, it is also computed during each of the other high level analysis steps.

In the high level analysis input dialog (Figure 11.12) the Factor assignment, chosen at the time of loading the CEL file archive, is displayed again. For high level analysis it is necessary to choose the factor that will be used for the fold change calculation.

The output of the high level analysis is a table containing the Affy IDs, fold change and p-values that are calculated based on the input parameters and factor-level assignment (Figure 11.13). This output can be also converted to an ExPlain gene set via the menu option "Gene set - Convert selected rows to gene set" of the analysis specific menu.

11.4.1 Meta-analysis

In addition to the four high level analysis methods that can be used for studying single data sets, a procedure called Rank Product which enables analyses of data from different origins, e.g. different laboratories, is provided within ExPlain. The procedure employs the RankProd algorithm developed by Hong and Breitling [[32]], which is based on the RankProduct method of Breitling et al. 2004 [[33]]. A brief description of the procedure is provided at the end of this section of the manual.

The Rank Product input dialog (Figure 11.14) requires that you specify at least one data source, plus that you select two or more columns for both the control (baseline) and the experiment. The selection can be made based on previously specified factors and levels by choosing the corresponding factor name in the selection option (factor) drop-down menu. Nota that up to three different sources can be processed by the procedure. Please also note that only data with affymetrix ID's present in all sets will be considered in the calculation. Therefore the data sources should show significant overlap in the probe sets for which expression data is provided.

The output of the Rank Product high level analysis will be a table containing Affy IDs as well as

Figure 11.11 RNA Degradation plot for a data set that contains data from arrays with high RNA degradation.

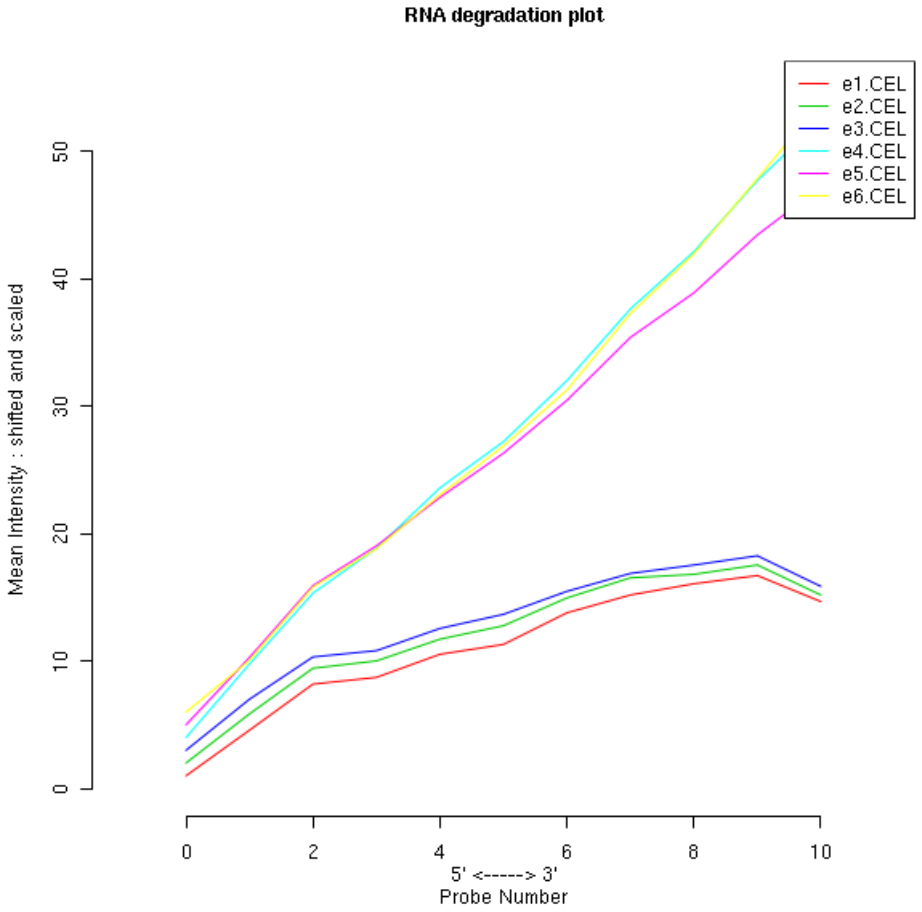
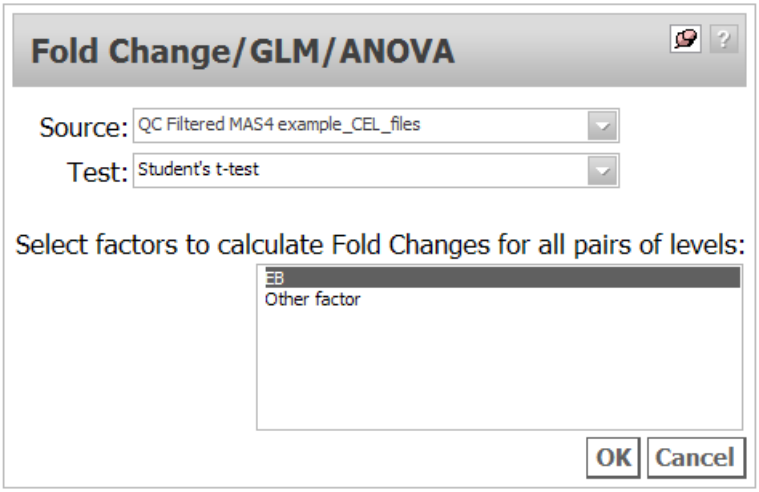


Figure 11.12 Fold Change - High level analysis dialog window

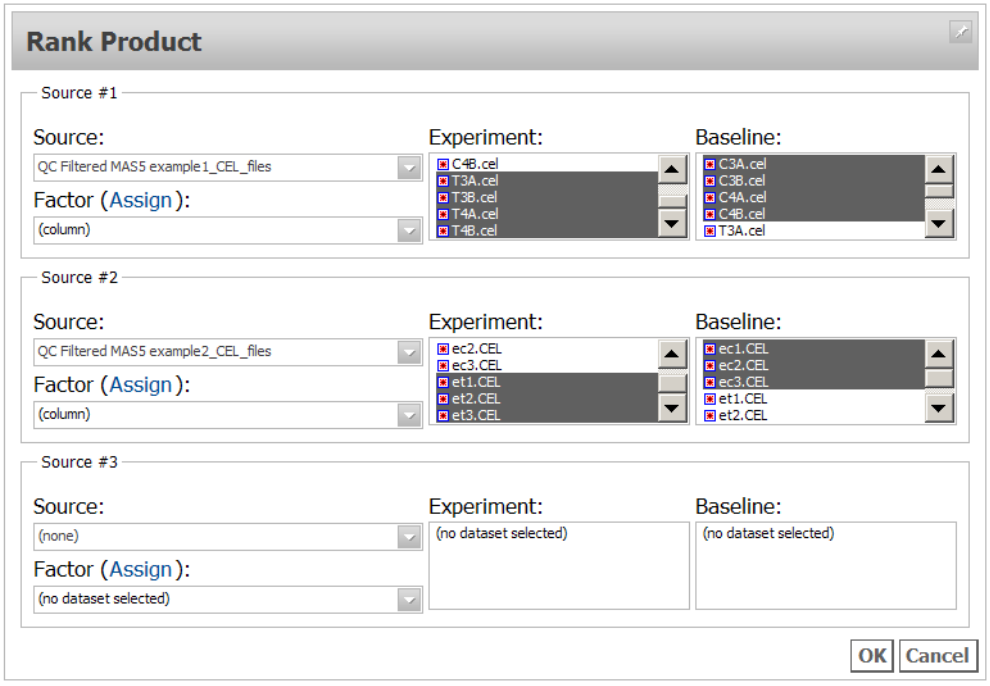


fold change-, rank product-, p- and fdr-values that are calculated based on the input parameters (Figure 11.15). This output can also be converted to an ExPlain gene set via the ”Gene set - Convert selected rows to gene set” of the analysis specific menu.

Figure 11.13 Output of High Level Analysis

	Affy ID ▲	log2 FC EB: E/B	P-value EB: E/B
☐	1007_s_at	-0.141016	0.0785341
☐	1053_at	-0.820644	0.000990837
☐	117_at	-0.555409	0.426196
☐	121_at	0.194605	0.0288328
☐	1255_g_at	0.183595	0.0450886
☐	1294_at	-0.522937	0.00419631
☐	1316_at	0.188997	0.433148
☐	1320_at	0.191603	0.0900062
☐	1405_i_at	0.225387	0.113561
☐	1431_at	1.21762	0.223415
☐	1438_at	0.0853689	0.444388
☐	1487_at	-0.429759	0.0032689
☐	1494_f_at	0.350762	0.0833852
☐	1598_g_at	-0.127858	0.00203457

Figure 11.14 RankProduct dialog window



11.4.2 Statistical Analyses of the gene expression data

The statistical analysis techniques (ANOVA, Fold change, Empirical Bayes and Generalized Linear Model) can be applied to a gene set, as well as to filtered CEL data. The "Statistical Analysis -> Fold Change" menu option of the "Analyze" menu provides a list of analysis methods. The dialog is displayed in Figure 11.16 below.

Select a gene set, method parameters, and a factor for the Fold Change computation. Only gene sets with the assigned factors can be used in this analysis. It is possible to change assignment from within the dialog by clicking the [Change configuration] button, or by launching the dialog from the "Data" menu.

Figure 11.15 Result of RankProduct analysis

[+] **Rank Product:** RP QC Filtered MAS5 example2 CEL files

Filter (filter bar): none (total 22283 rows)

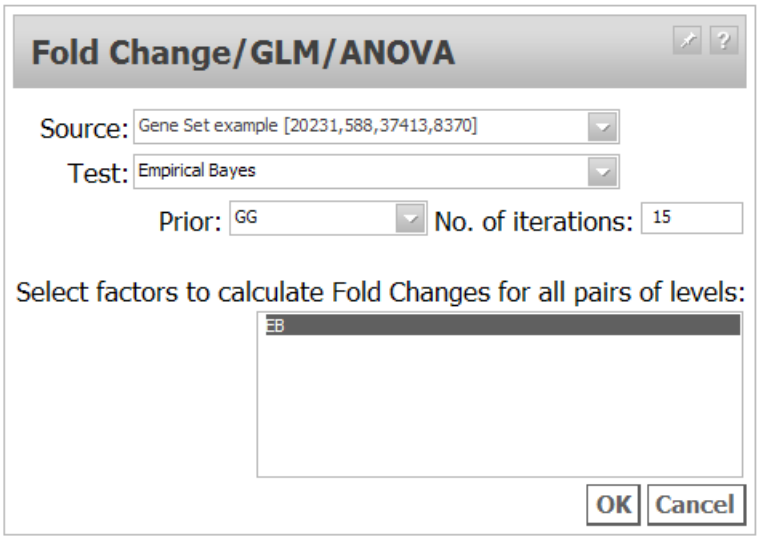
Export: Plain text | XLS | RTF

Mark: Page (100) | All (22283) | None | Invert

Affy ID	Fold Change	RankProduct	p-value	fdr
1007_s_at	0.75107	16740.8	0.9968	1.1104
1053_at	0.878604	14381.4	0.9682	1.1941
117_at	1.22402	3928.24	0.0413	0.4038
121_at	1.03367	8940.48	0.5416	1.0616
1255_g_at	1.35247	6463.14	0.2352	0.8031
1294_at	0.787509	17071.1	0.9978	1.0978
1316_at	1.13516	6702.36	0.2618	0.83
1320_at	1.093	8139.44	0.4388	0.9919
1405_i_at	1.2403	5565.86	0.1465	0.6887

Note that some analyses require at least two columns for each level.

Figure 11.16 Statistical analysis dialog



A resulting gene set contains genes from the original set and columns with the statistical data calculated. A fragment of the result of the "Empirical Bayes" analysis is shown in Figure 11.17.

The Rank Product procedure described in the CEL file high level analysis section can also be applied to up to three gene sets, provided there are columns with absolute expression values for controls and experiments assigned to the set. The input dialog window is identical to the one shown in Figure 11.14.

The resulting gene set contains the genes present in all source sets and several columns with statistical data calculated. A fragment of the result of the "Rank Product" analysis is shown in Figure 11.18.

Figure 11.17 The Empirical Bayes result

Gene symbol	BKL description	log2 FC EB: E/B	(B).ne.E	(E=B)	1-(E=B)
CCDC11	Protein of unknown function; has high similarity to uncharacterized mouse Ccdc11	4.83588	1	4.16777e-13	1
HHATL	Protein with low similarity to human HHAT; which binds to GTP and may be associated with melanoma	4.74761	1	1.14984e-12	1
PPP1R10	Protein phosphatase 1 regulatory (inhibitor) subunit 10; increases hypoxia-induced cell death via post translational modification of TP53 and MDM2; induces mitotic chromosome decondensation; gene is aberrantly expressed in Alzheimer disease	4.47569	1	1.23481e-11	1
AMH	Anti-Mullerian hormone; plays a role in the induction of apoptosis; increased expression is associated with polycystic ovary syndrome; obesity; and hypothalamic amenorrhea; mutation in the gene is associated with persistent Mullerian duct syndrome	4.4127	1	3.80546e-11	1
ZFP2	Protein with strong similarity to zinc finger protein 2 (mouse Zfp2); which binds zinc and is a putative transcription factor that may be required for neuronal development; contains 13 zinc finger C2H2 type repeats; which bind nucleic acids	3.99162	1	2.08193e-08	1
KCND1	Potassium voltage-gated channel Shal-related subfamily member 1; a transmembrane transporter that is involved in potassium ion transport and response to peptide hormone stimulus; may act in cell volume homeostasis and regulation of heart contraction	3.78529	1	2.26904e-08	1
PAGE2	Protein of unknown function; has strong similarity to uncharacterized human PAGE2B	3.73435	1	4.66172e-07	1
LRRC14	Protein containing three leucine rich repeats; which mediate protein-protein interactions; has low similarity to preferentially expressed antigen of melanoma (human PRAME); which is a tumor antigen overexpressed in leukemias	3.72687	1	1.78645e-08	1

Figure 11.18 "Rank Product" analysis result

Gene symbol	Description	Fold Change	RankProduct	p-value	fdr
A1BG	alpha-1-B glycoprotein	0.835378	7317.66	0.6586	1.156
A1CF	APOBEC1 complementation factor	0.958689	6569.4	0.5347	1.1283
A2BP1	ataxin 2-binding protein 1	1.04371	6554.83	0.5321	1.1265
A2M	alpha-2-macroglobulin	0.580866	10622	0.967	1.0575
A2ML1	alpha-2-macroglobulin-like 1	0.885263	9403.71	0.9027	1.103
A4GALT	alpha 1,4-galactosyltransferase	0.841252	5055.62	0.2741	0.952
A4GNT	alpha-1,4-N-acetylglucosaminyltransferase	0.936106	6299.55	0.4878	1.1107
AAAS	achalasia, adrenocortical insufficiency, alacrimia (Allgrove, triple-A)	2.46028	1444.53	0.0015	0.09
AACS	acetoacetyl-CoA synthetase	1.05135	7420.5	0.6745	1.1602
AADAC	arylacetamide deacetylase (esterase)	0.648968	10739.5	0.9707	1.0535
AADACL1	arylacetamide deacetylase-like 1	1.22695	2754.89	0.0304	0.4088

11.4.3 The Rank Product algorithm

The Rank Product procedure was developed by Hong and Breitling [[32]], based on the RankProduct method of Breitling et al. 2004 [[33]]. This procedure allows a detection of differently expressed genes by ranking the genes according to the difference in expression between control and experiment. After calculating the expression_expreriment / expression_control ratio for all possible comparisons (k), the genes are sorted according to the fold change ratios determined for each comparison in k separate lists. The occurence of a gene (g) being differently expressed is evaluated by calculating its rank product (RP) regarding the ranking positions of the genes in all sorted lists.

quation 11.4.1 Rank Product

$$RP_g = \left(\prod_{i=1}^{k} r_{i,g} \right)^{\frac{1}{k}}$$

RP_g = RankProduct of gene g
r_{ig} = rank of gene g in comparison list i
k = total number of comparisons

To estimate the probability of observing a specific RP value, a p-value is determined in analogy to the e-value of the BLAST procedure [[34]]. First the RP values of a large number of random permutations of the original expression values are calculated. Then these values are used to evaluate the likeliness of observing a RP value equal to or smaller than the one under consideration by random chance. Additionally the percentage of false positives or false discovery rate, when all genes with RP values smaller than the RP under consideration are regarded as differently expressed, is estimated based on the permuted values.

11.5 CRC clustering

The weighted Chinese restaurant clustering algorithm developed by Zhaohui S. Qin [[35]], enables clustering of genes according to their expression values. This model-based clustering procedure has the advantage that unlike many other clustering algorithms the number of clusters is determined during cluster assignment, so that the expected number of clusters does not need to be specified in advance.

The algorithm requires as input a gene set with expression data of different experiments (Figure 11.19), e.g. different treatments or a time series study. Expression values of biological replicas should be averaged before the clustering process is started. Average expression values can be calculated using the option "Data -> Calculate from existing columns" in the ExPlain top menu.

Figure 11.19 Example of a data set that can be analyzed by CRC clustering.

	accession	Species	2 hours	4 hours	6 hours	8 hours	10 hours	12 hours	16 hours	18 hours	24 hours
(*)	PLAC8	Human	-0.0737	-0.0556382	-0.157988	-0.160999	-0.100793	-0.0014528	0.0406914	0.112939	0.194217
(*)	SAP30	Human	-0.0652601	-0.158579	-0.254909	-0.233837	-0.140518	0.0069871	-0.0170953	0.0280592	-0.0803116
(*)	RANBP9	Human	-0.0636394	-0.147928	-0.158464	-0.147928	-0.0290209	0.0086078	0.0522572	0.124504	0.0703189
(*)	FIG4	Human	-0.102537	-0.0935058	-0.253052	-0.228969	-0.289175	-0.0212586	0.0359371	0.132267	0.0991534
(*)	PRKCZ	Human	-0.0040509	-0.085329	-0.0793084	0.0621757	0.0712066	0.173557	0.20968	0.185598	0.170546
(*)	ZMYND8	Human	-0.214918	-0.413598	-0.455742	-0.323288	-0.205887	0.0198857	0.083102	0.0981535	0.0981535
(*)	NFATC1	Human	-0.224882	-0.261006	-0.285088	-0.0803876	-0.10748	0.124313	0.0490552	0.0039008	0.0039008
(*)	SMAD4	Human	-0.0751548	0.0663293	0.0693396	0.0422469	0.102453	0.147607	0.102453	0.132556	0.0753602
(*)	OTUB1	Human	0.115798	0.0977358	0.121818	0.112787	0.0977358	-0.0377277	-0.0046144	-0.0557895	-0.0347174
(*)	TM2D1	Human	-0.144069	-0.150089	-0.183203	-0.246419	-0.107945	-0.0597803	-0.086873	0.0214978	0.0094566
(*)	RTF1	Human	-0.0502086	-0.113425	-0.0622498	-0.0291365	0.0340798	0.151481	0.0792343	0.121379	0.0250489
(*)	TREX1	Human	0.110068	0.158233	0.200377	0.189841	0.0769545	-0.127746	-0.0524884	-0.0600141	-0.0916223

CRC clustering can be carried out by clicking on "CRC clustering" in the "Analyze" drop down menu. In the input form (Figure 11.20) you must specify the parameters that shall be used for clustering. If your set contains a large number of genes (more than 3000 genes) it is strongly recommended that you perform a filtering step before clustering. Filtering is performed using a slightly modified form of the procedure employed by Tamayo et al. [[36]]. To filter the gene set, mark the checkbox "Use most differentially expressed genes" and specify the number of genes that shall be present after filtering. Genes passing the filtering step are those with a high difference in expression between experiments, i.e. these genes posses pronounced expression profiles. The exact calculation procedure employed in the filtering step as well as details about CRC clustering can be found in the section "Algorithm details of CRC clustering".

Parameters for the probability threshold and the maximal shift size must be specified. During the clustering process the probability that a gene actually belongs to the assigned cluster is calculated. Probability values close to zero denote that the genes are only weakly associated with the assigned cluster, while values between 0.9 and 1 mean that it is very likely that the genes belong to the assigned cluster. If a posterior probability threshold is specified, all genes with probability values below this threshold will be excluded from the clustering results.

In adition of clustering genes exhibiting a similar expression profile, The CRC clustering algorithm can also identify complex correlation relationships like time shift effects. If you want to find genes that show a similar, but time shifted, profile please specify the number of shifts (0, 1, 2 ...) that will be regarded in the analysis. If for example a shift size of 3 is used, profiles $(x_{i1}, x_{i2}, \ldots, x_{iM-2})$, $(x_{i2}, x_{i3}, \ldots, x_{iM-1})$ and $(x_{i3}, x_{i4}, \ldots, x_{iM})$ of gene i will be compared with each of the existing clusters to determine if there is a cluster that fits one of the profiles well.

When the clustering process is completed, the clusters appear as separate gene sets in the ExPlain tree (Figure 11.21). Beside the name of the cluster and the parameters used for CRC two values are given in brackets in the tree entry. The first value is the cluster tightness, which represents a measure of the homogeneity of the genes within the cluster. This value is followed by the cluster stability measure, which reflects the similarity of the genes in the cluster. A high stability value means that the genes within the cluster are closely related. The generated gene sets contain an additional column in which the posterior probability calculated for each gene is listed, so that the genes belonging to a cluster can be filtered by their probability of belonging to the cluster after the clustering process has been completed.

Figure 11.20 CRC Input dialog window

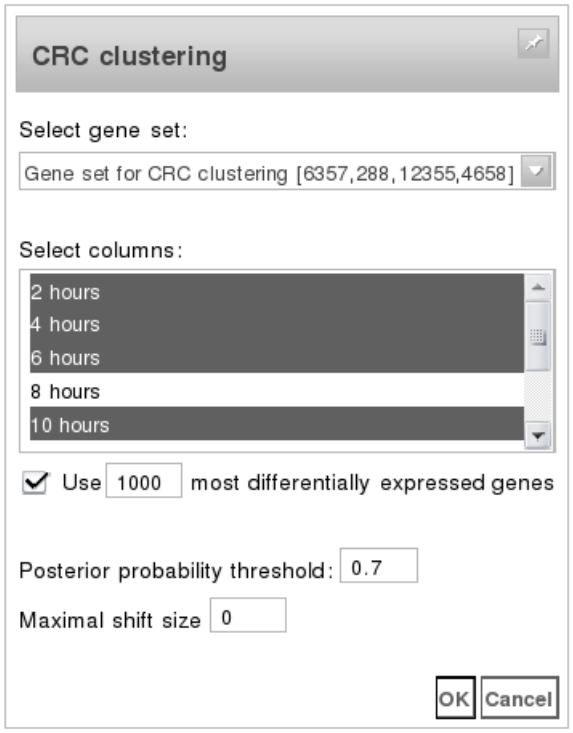
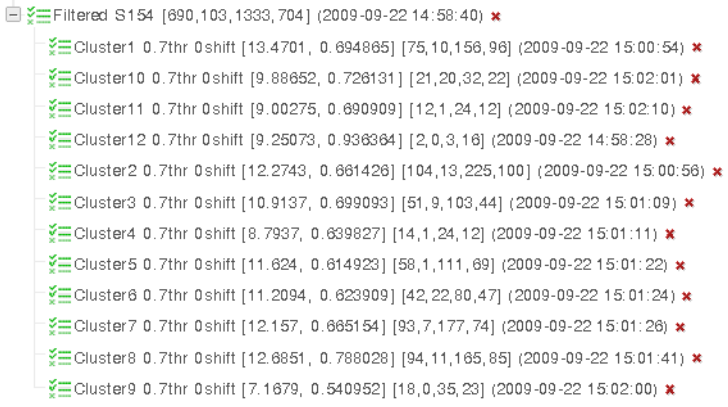


Figure 11.21 Section of the results of a CRC clustering analysis displayed in the ExPlain tree.



11.5.1 Algorithm details of CRC clustering

As recommended by Qin 2006 [[37]] genes showing little variation across all experiments can be removed by filtering the data set before the clustering procedure is started. In the filtering step the minimum and maximum expression values are determined for each gene regarding all experiments. Then relative (rel) and absolute (abs) difference values are calculated according to the formulas:

$$rel = max/min \qquad abs = max\text{-}min \,/\, max\text{+}min$$

Subsequently all genes for which abs > 1.1 are ranked according to their relative difference values, and the X top ranked genes are used as input set for CRC.

CRC is a model-based clustering approach based on the Chinese restaurant process (CRP) [[38]]. The procedure is analogous to the random seating of customers that sequentially arrive at a restaurant with an infinite number of tables and an unlimited number of chairs. Each new customer that arrives will be seated according to the actual seating scheme of the persons that have arrived at the restaurant before. Translated to the clustering of gene expression data this means that each gene is assigned to a cluster depending on the existing constellation of gene to clusters assignment. The number of clusters and the parameters are determined in the course of the process, so that no initial information about the clusters

is needed.

The basic clustering procedure consists of the following steps:
(1) Initialization: Genes are randomly assigned to a haphazardly number of clusters.
(2) Re-assignment: Each gene is again assigned to a cluster by:

a) Removal of the gene from its actual cluster and calculation of the probability that the gene joins any of the presently existing clusters or starts a new cluster, taking the current assignment of all other genes into account.

b) Assignment of the gene to a cluster according to probability values, and update of variables, e.g. number of clusters.

Steps a) and b) of (2) are carried out for each gene and repeated for a large number of cycles until convergence is reached.

Details about the probability value calculation can be found in the article of Qin 2006 [[35]].

In order to obtain information about how tightly the genes are associated with a certain cluster the Bayes ratio of the clusters is calculated. For each cluster two likelihood values are determined, the likelihood that all genes in the cluster follow the same set of normal distributions, and the likelihood that each gene follows a set of normal distributions that differs from that of all other genes. Suppose that the expression levels of N genes from M experiments are collected. The expression data can be denoted as:

$x = \{x_{ij}, i=1,\ldots,N, j = 1,\ldots,M\}$

Under the assumption that the first n1 genes belong to the same cluster the Bayes ratio can be calculated by:

quation 11.5.1 Bayes ratio

$$\frac{\prod_{j=1}^{M}\int P\left(x_{1j}, x_{2j},\ldots,x_{n_1 j} \mid \mu_j \sigma_j^2\right) P\left(\mu_j, \sigma_j^2\right) d\mu_j \, d\sigma_j^2}{\prod_{j=1}^{M}\prod_{i=1}^{n_1}\int P\left(x_{ij} \mid \mu_{ij}, \sigma_{ij}^2\right) P\left(\mu_{ij}, \sigma_{ij}^2\right) d\mu_{ij} \, d\sigma_{ij}^2}$$

The cluster tightness given in the ExPlain tree entries is the log Bayes ratio normalized by the number of genes in the respective cluster.

Cluster stability is measured using the similarity measure defined by Medvedovic and Sivaganesan 2000 [[39]]. The similarity measure between a pair of genes is equivalent to the proportion of times during iteration in which both genes are assigned to the same cluster. To obtain the stability of a cluster, the average of the similarity measures of all gene pairs within the cluster is calculated.

Table 11.2

Background correction	Normalize method	PM correction method	Summarization method
MAS	Constant	PM Only	Liwong
None	Invariantset	Subtract MM	MAS
RMA	Quantiles		Median Polish
	Contrasts		Avgdiff
	LOESS		Playerout
	Qspline		
	Qspline.Robust		

Chapter 12

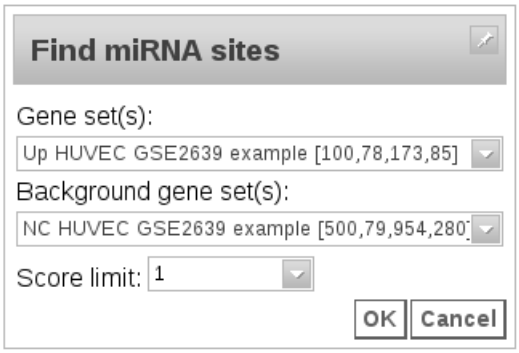
miRNA analysis

12.1 Identification of miRNA targets

This feature identifies putative miRNA binding sites in the upstream regions of genes based on TargetScan [[41]] data. Pre-calculated total context scores were taken from TargetScan <http://www.targetscan.org/cgi-bin/targetscan/data_download.cgi?db=vert_50>. Lower values correspond to sites of better quality.

Use the "miRNA sites" option within the "Analyze" menu to launch the miRNA search dialog.

Figure 12.1 Find miRNA sites dialog window



First specify the query and background gene sets. Note that you can choose "none" at the top of the background menu if you wish to identify miRNA binding sites without comparing to a background set.

Next choose a score limit. If the total context score of predicted sites for a given miRNA on a given gene is greater than the selected limit, then these sites will not be considered. As the total context score doesn't exceed 1, a value of 1 in this field will turn this filter off. Pressing the OK button, you will start the miRNA target identification process. The output result is displayed below.

The result of this search contains a list of miRNA names, their descriptions and lists of identified target genes, followed by the count of target genes, the total number of sites found, the average score and the sum score.

The average score represents the average value of the total context score values for all genes, while the sum score represents the sum of total context score values for all genes.

The miRNA search result node opens specific options within the "miRNA analysis results" menu item which allow you to extract new subsets of selected miRNA genes, or their targets.

The P-value is calculated based on hypergeometric distribution using the total and matched gene counts in the data set and background set.

Figure 12.2 miRNA search output

[+] **MiRNA analysis result:** miRNA 1 Up HUVEC GSE2639 example

Filter (filter bar): none (total 531 rows)

Rows per page: 100

Export: Plain text | XLS | RTF

1 2 3 4 5 6

Mark: Page (100) | All (531) | None | Invert

	MiRNA name	MiRNA description	Genes list	Genes count	Genes count in background	Total sites	Average score	Sum score	Genes count ratio	P-value ▲
(*)	MIR944	microRNA 944	ATP1B1, BHLHE40, BMP2, C8orf4, CAB39, CD47, CSF2, CXCL1, CXCL5, F2RL1, GATA6, GBP1, HLA-B, IL18R1, IL8, JAG1, MMP10, NAV2, NRIP1, OPTN, PLAU, PMAIP1, PTGS2, PTPRK, RAPGEF5, SELE, SERPINB2, SLC12A2, SLC7A2, SMAD3, SOD2, TNFAIP1, TNFAIP2, TNFRSF9, TRAF1, VCAM1, VEGFA	37	90	50	-0.374876	-13.8704	2.05556	4.49461e-05
(*)	MIR548C	microRNA 548c	ABCG1, BHLHE40, BIRC3, C11orf17, C8orf4, CAB39, CD47, CD69, CSF2, CXCL2, CXCL3, CXCL5, CXCR7, DDX58, ETS1, F3, FZD6, GATA6, GBP1, HIVEP2, ICAM1, IER5, IFNGR1, IL18R1, IL8, ISG20, JAG1, LAMC2, NAV2, NFKBIA, NRIP1, NUAK2, OPTN, PFKFB3, PLAU, PMAIP1, PPAP2B, PTGS2, PTPRK, RAPGEF5, RIPK2, SDC4, SELE, SERPINB2, SLC12A2, SLC31A2, SLC7A2, SMAD3, SSH1, TNFAIP2, ...	57	179	108	-0.402567	-22.9463	1.59218	6.98384e-05
(*)	MIR374A	microRNA 374a	BHLHE40, BIRC3, BMP2, CAB39, CCL2, CCL20, CD47, CD69, CXCL5, F2RL1, FZD6, GATA6, GBP1, HIVEP2, HLA-B, IL18R1, PANX1, PMAIP1, PPAP2B, PTGS2, SELE, SLC12A2, TNFAIP3, TNFRSF9, UGCG, VEGFA	26	56	37	-0.413173	-10.7425	2.32143	0.000204694
(*)	MIR942	microRNA 942	ABCG1, BHLHE40, CD47, CD69, CXCL5, DDX58, EFNA1, ETS1, F2RL1, GATA6, GFPT2, HIVEP2, IL18R1, IL7R, IRF1, NAV2, NFKBIE, OPTN, PFKFB3, PPAP2B, RAPGEF5, SDC4, SELE, SLC7A2, SMAD3, SSH1, TNFSF10, TRAF1, UGCG, VEGFA	30	74	39	-0.200407	-6.0122	2.02703	0.000411583

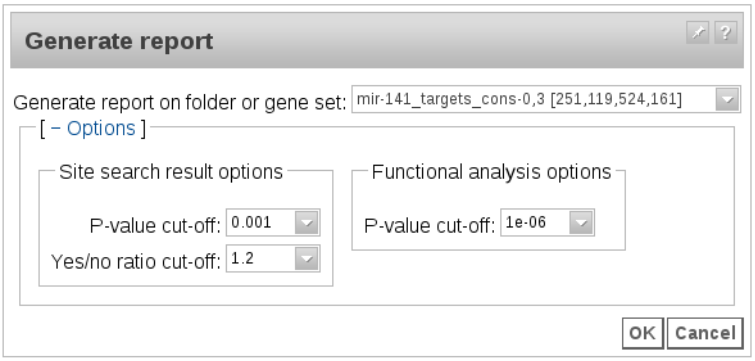
Chapter 13

Reports

13.1 Report generation

ExPlain provides the ability to create a brief report on your data and analysis results. The report is available for the folder and for the gene set. Short descriptions of all the data, analysis performed and main results are saved in a report tree node for the selected folder or gene set. Use the "Generate report" link within the "Analyze" menu to launch the report creation dialog.

Figure 13.1 Simple report generation dialog window



Select the desired folder or gene set, depending on whether you want to collect information on the whole folder or on a gene set with its subnodes. In advanced dialog mode you can also set up the threshold values to restrict the output results. Pressing the OK button, initiates the report generation. The output result is displayed below.

13.2 Graph report generation

The graph report is a tree item type which contains several graphs.

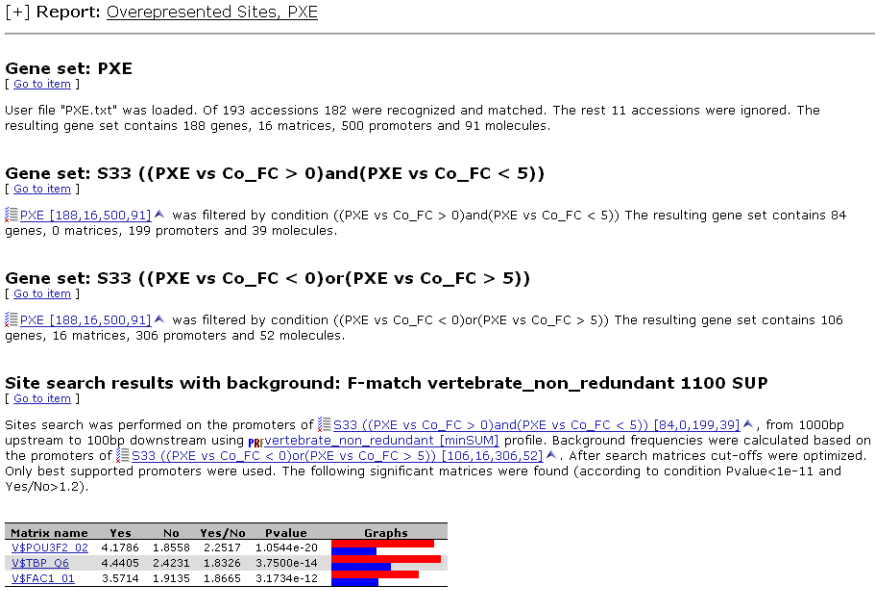
13.2.1 Generating a graph of the value distribution of a gene set

You can create a graphical representation of up-regulated, down-regulated or non-differentially expressed genes. To do so, choose the "Distribution of values" option from the "Analyze" menu, select a gene set to be used, specify the objects to consider and an expression column, and press the OK button (Figure 13.3). A graph will appear under the original gene set in the tree. In addition, you have the option to export this graph as an RTF document.

13.2.2 Graphical report on intervals

Using the "Graph report on interval" feature you can generate graphs on your interval set displaying the distribution of intervals by distance from the TSS.

Figure 13.2 Generated report



[+] Report: Overepresented Sites, PXE

Gene set: PXE
[Go to item]

User file "PXE.txt" was loaded. Of 193 accessions 182 were recognized and matched. The rest 11 accessions were ignored. The resulting gene set contains 188 genes, 16 matrices, 500 promoters and 91 molecules.

Gene set: S33 ((PXE vs Co_FC > 0)and(PXE vs Co_FC < 5))
[Go to item]

PXE [188,16,500,91] was filtered by condition ((PXE vs Co_FC > 0)and(PXE vs Co_FC < 5)) The resulting gene set contains 84 genes, 0 matrices, 199 promoters and 39 molecules.

Gene set: S33 ((PXE vs Co_FC < 0)or(PXE vs Co_FC > 5))
[Go to item]

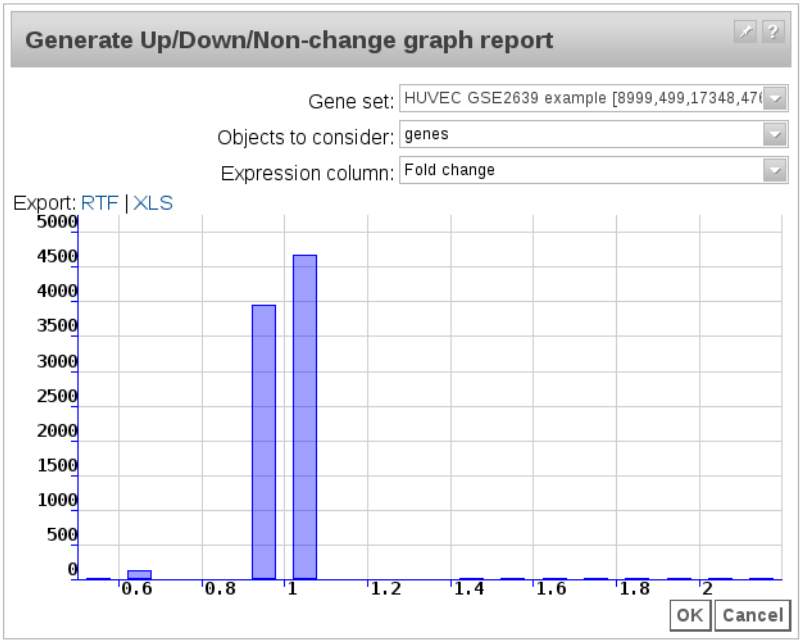
PXE [188,16,500,91] was filtered by condition ((PXE vs Co_FC < 0)or(PXE vs Co_FC > 5)) The resulting gene set contains 106 genes, 16 matrices, 306 promoters and 52 molecules.

Site search results with background: F-match vertebrate_non_redundant 1100 SUP
[Go to item]

Sites search was performed on the promoters of S33 ((PXE vs Co_FC > 0)and(PXE vs Co_FC < 5)) [84,0,199,39], from 1000bp upstream to 100bp downstream using vertebrate_non_redundant [minSUM] profile. Background frequencies were calculated based on the promoters of S33 ((PXE vs Co_FC < 0)or(PXE vs Co_FC > 5)) [106,16,306,52]. After search matrices cut-offs were optimized. Only best supported promoters were used. The following significant matrices were found (according to condition Pvalue<1e-11 and Yes/No>1.2).

Matrix name	Yes	No	Yes/No	Pvalue	Graphs
V$POU3F2_02	4.1786	1.8558	2.2517	1.0544e-20	
V$TBP_Q6	4.4405	2.4231	1.8326	3.7500e-14	
V$FAC1_01	3.5714	1.9135	1.8665	3.1734e-12	

Figure 13.3 Generating graph on Up-/Down-regulated/Non-differentially-expressed genes



Launch the graph report dialog using either the "Generate graph on interval" link from the "Analyze" menu, or the toolbar button, or the link from "Interval" menu when standing on an interval node in the process tree.

As shown in the dialog depicted in Figure 13.4 you may specify the interval set you want to analyze and optionally specify a gene set to filter your interval set. When specifying this gene set, only those promoters from the interval set that are present in gene set will be used to plot the graphs. The "Blur window" option specifies the window width in which values will be averaged to create a smoother graph. When 'Display original distribution' is checked, non-smoothed graph will be displayed, which is equivalent to blur window = 0 bp.

The result is a graph report tree item (Figure 13.5) displaying the distribution of interval counts and signal/p-value distributions, if this information is attached to your interval set.

Figure 13.4 Graph report dialog

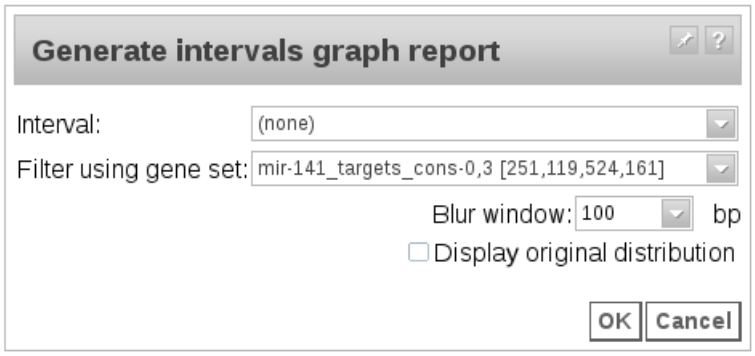
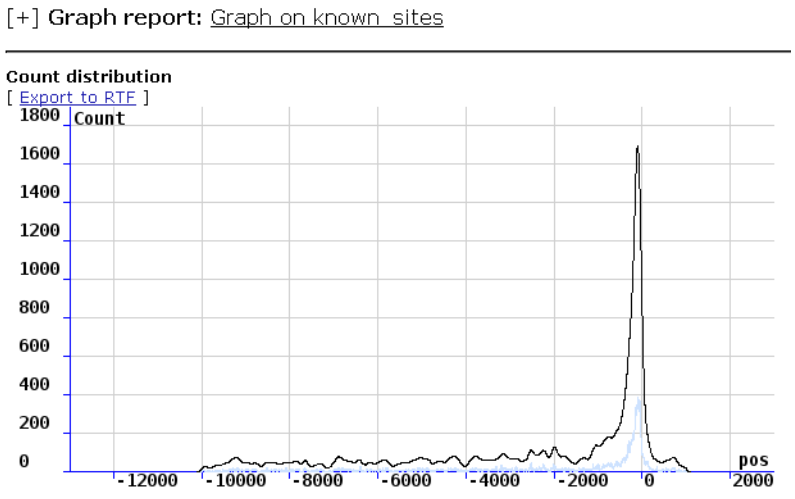


Figure 13.5 Graph report



Part II

Appendix

Chapter 14

References

1 **TRANSFAC and TRANSCompel.**

Matys V., Kel-Margoulis O.V., Fricke E., Liebich I., Land S., Barre-Dirrie A.,

Reuter I., Chekmenev D., Krull M., Hornischer K., Voss N., Stegmaier P., Lewicki-Potapov B.,

Saxel H., Kel A.E. and Wingender E. TRANSFAC™ and its module TRANSCompel™: transcriptional

gene regulation in eukaryotes Nucleic Acids Res. 34, D108-D110 (2006).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=16381825>

TRANSFAC™ <http://www.biobase-international.com/pages/index.php?id=40>

TRANSCompel™ <http://www.biobase-international.com/pages/index.php?id=112>

2 **Composite Module Analyst (CMA).**

Kel A., Konovalova, T., Valeev, T., Cheremushkin, E., Kel-Margoulis, O. and Wingender, E.

Composite Module Analyst: a fitness-based tool for identification of transcription factor binding site combinations

Bioinformatics 22, 1190-1197 (2006).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=16473870>

CMA <http://www.gene-regulation.com/pub/programs.html#CMAnalyst>

3 **TRANSPATH.**

Krull M., Pistor S., Voss N., Kel A.E., Reuter I., Kronenberg D., Michael H., Schwarzer K., Potapov A., Choi C., Kel-Margoulis O.V. and Wingender E.

TRANSPATH™: An Information Resource for Storing and Visualizing Signaling Pathways and their Pathological Aberrations

Nucleic Acids Res. 34, D546-D551 (2006).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=16381929>

TRANSPATH™ <http://www.biobase-international.com/pages/index.php?id=39>

4 **CYTOMER.**

Michael H., Chen X., Fricke E., Haubrock M., Ricanek R. and Wingender, E.

Deriving an ontology for human gene expression sources from the CYTOMER™ database on human organs and cell types

In Silico Biol. 5, 0007 (2004).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15972006>

CYTOMER™ <http://www.gene-regulation.com/pub/databases/cytomer/>

5 **Human Protein Survey Database (HumanPSD).**

Hodges, P.E., Carrico, P.M., Hogan, J.D., O'Neill, K.E., Owen, J.J., Mangan, M., Davis, B.P., Brooks, J.E. and Garrels, J.I.

Annotating the human proteome: the Human Proteome Survey Database (HumanPSD) and an in-depth target database

for G protein-coupled receptors (GPCR-PD) from Incyte Genomics

Nucleic Acids Res. 30, 137-141 (2002).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=11752275>

HumanPSD™ <http://www.biobase-international.com/pages/index.php?id=71>

6 **TRANSPro.**

Chen X., Wu J.-m., Hornischer K., Kel A. and Wingender E.

TiProD: The Tissue-specific Promoter Database

Nucleic Acids Res. 34, D104-D107 (2006).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=16381824> TRANSPro™ <http://www.biobase-international.com/pages/index.php?id=113>

7 **MATCH.**

Kel, A.E., Goessling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., Wingender, E.

MATCH™: a tool for searching transcription factor binding sites in DNA sequences

Nucleic Acids Res. 31, 3576-3579 (2003).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=12824369>

8 **Source of the human housekeeping gene set.**

Eisenberg E. and Levanon E.Y.

Human housekeeping genes are compact

Trends Genet. 19, 362-365 (2003).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=12850439>

9 **RefSeq.**

Pruitt K.D., Tatusova T. and Maglott D.R.

NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins

Nucleic Acids Res. 33, D501-D504 (2005).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15608248>

RefSeq <http://www.ncbi.nlm.nih.gov/RefSeq/>

10 **Human Gene Nomenclature Database (HGNC).**

Wain H.M., Lush M.J., Ducluzeau F., Khodiyar V.K. and Povey S.

Genew: the Human Gene Nomenclature Database, 2004 updates

Nucleic Acids Res. 32, D255-D257 (2004)

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=14681406>

HGNC <http://www.gene.ucl.ac.uk/nomenclature/index.html>

11 **Mouse Genome Informatics (MGI).**

Eppig J.T., Bult C.J., Kadin J.A., Richardson J.E., Blake J.A. and the members of the Mouse Genome Database Group

The Mouse Genome Database (MGD): from genes to mice - a community resource for mouse biology

Nucleic Acids Res. 33, D471-D475 (2005).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15608240>

MGI <http://www.informatics.jax.org/>

12 **Rat Gene Nomenclature Comittee (RGNC).**

RGNC <http://rgnc.gen.gu.se/RGNChem.html>

13 **EMBL.**

Cochrane G., Aldebert P., Althorpe N., Andersson M., Baker W., Baldwin A., Bates K., Bhattacharyya S., Browne P.,

Van Den Broek A., Castro M., Duggan K., Eberhardt R., Faruque N., Gamble J., Kanz C., Kulikova T., Lee C.,

Leinonen R., Lin Q., Lombard V., Lopez R., McHale M., McWilliam H., Mukherjee G., Nardone F., Pastor M.P., Sobhany S.,

Stoehr P., Tzouvara K., Vaughan R., Wu D., Zhu W. and Apweiler R.

EMBL Nucleotide Sequence Database: developments in 2005

Nucleic Acids Res. 34, D10-D15 (2006).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=16381823>

EMBL <http://www.ebi.ac.uk/embl/>

14 **GenBank.**

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L.

GenBank

Nucleic Acids Res. 33, D34-D38 (2005).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15608212>

GenBank <http://www.ncbi.nih.gov/Genbank/>

15 **DDBJ.**

Okubo K., Sugawara H., Gojobori T. and Tateno Y.

DDBJ in preparation for overview of research activities behind data submissions

Nucleic Acids Res. 34, D6-D9 (2006).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=16381940>

DDBJ <http://www.ddbj.nig.ac.jp/>

16 **Ensembl.**

Hubbard T., Andrews D., Caccamo M., Cameron G., Chen Y., Clamp M., Clarke L., Coates G., Cox T., Cunningham F.,

Curwen V., Cutts T., Down T., Durbin R., Fernandez-Suarez X.M., Gilbert J., Hammond M., Herrero J., Hotz H.,

Howe K., Iyer V., Jekosch K., Kahari A., Kasprzyk A., Keefe D., Keenan S., Kokocinsci F., London D., Longden I.,

McVicker G., Melsopp C., Meidl P., Potter S., Proctor G., Rae M., Rios D., Schuster M., Searle S., Severin J.,

Slater G., Smedley D., Smith J., Spooner W., Stabenau A., Stalker J., Storey R., Trevanion S., Ureta-Vidal A.,

Vogel J., White S., Woodwark C. and Birney E.

Ensembl 2005

Nucleic Acids Res. 33, D447-D453 (2005).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15608235>

Ensembl <http://www.ensembl.org/index.html>

17 **Entrez Gene.**

Maglott D., Ostell J., Pruitt K.D. and Tatusova T.

Entrez Gene: gene-centered information at NCBI

Nucleic Acids Res. 33, D54-D58 (2005).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15608257>

Entrez Gene <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=gene>

18 **Entrez Protein.**

Wheeler D.L., Church D.M., Federhen S., Lash A.E., Madden T.L., Pontius J.U., Schuler G.D., Schriml L.M., Sequeira E., Tatusova T.A. and Wagner L.

Database Resources of the National Center for Biotechnology

Nucl Acids Res. 31, 28-33 (2003).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=12519941>

Entrez Protein <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=protein>

19 **International Protein Index (IPI).**

Kersey P.J., Duarte J., Williams A., Karavidopoulou Y., Birney E. and Apweiler R.

The International Protein Index: An integrated database for proteomics experiments

Proteomics 4, 1985-1988 (2004).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15221759>

IPI <http://www.ebi.ac.uk/IPI/index.html>

20 **UniGene.**

Wheeler D.L., Church D.M., Federhen S., Lash A.E., Madden T.L., Pontius J.U., Schuler G.D., Schriml L.M., Sequeira E., Tatusova T.A. and Wagner L.

Database Resources of the National Center for Biotechnology

Nucl Acids Res. 31, 28-33 (2003).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=12519941>

UniGene <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene>

21 **UniProt.**

Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N. and Yeh L.S.

The Universal Protein Resource (UniProt)

Nucleic Acids Res. 33, D154-D159 (2005).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15608167>

UniProt <http://www.uniprot.org/>

22 **Gene Ontology.**

The Gene Ontology Consortium

Gene Ontology: tool for the unification of biology

Nat. Genet. 25, 25-29 (2000).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=10802651>

Gene Ontology <http://www.geneontology.org>

23 **Benjamini/Hochberg multiple testing correction.**

Benjamini Y. and Hochberg Y.

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

Journal of the Royal Stat. Soc. 57, 289-300 (1995).

24 **Eukaryotic Promoter Database (EPD).**

Schmid C.D., Perier R., Praz V. and Bucher, P.

EPD in its twentieth year: towards complete promoter coverage of selected model organisms

Nucleic Acids Res. 34, D82-D85 (2006).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=16381980>

EPD <http://www.epd.isb-sib.ch/index.html>

25 **DBTSS.**

Suzuki Y., Yamashita R., Sugano S. and Nakai K.

DBTSS, DataBase of Transcriptional Start Sites: progress report 2004

Nucleic Acids Res. 32, D78-D81 (2004).

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=14681363>

DBTSS <http://dbtss.hgc.jp/>

26 **The Gene Set Enrichment Analysis.**

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander and Jill P. Mesirov

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.

PNAS 43, 15545-15550 (2005).

PNAS <http://www.pnas.org/cgi/content/full/102/43/15545>

27 **The Kolmogorov-Smirnov-like statistic.**

Hollander M. and Wolfe D.A. (1999) Nonparametric Statistical Methods (Wiley, New York).

28 **LibGD library.** <http://www.libgd.org/Main_Page>

29 **affyQCReport.**

Craig Parman and Conrad Halling (2005)

affyQCReport: A Package to Generate QC Reports for Affymetrix Array Data.

<http://www.bioconductor.org/packages/bioc/html/affyQCReport.html>.

30 **QC and Affymetrix data.**

Wilson C.,Pepper S. D. and Miller C. J. QC and Affymetrix data

Paterson Institute for Cancer Research, Christie Hospital NHS Trust, Manchester, UK.

31 **Simpleaffy.**

Crispin J. Miller (2005) simpleaffy.

<http://bioinformatics.picr.man.ac.uk/simpleaffy/index.jsp>.

32 **RankProd.**

Hong F., and Breitling R. (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics, 24, 374-82.

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=18204063>

33 **RankProduct.**

Breitling R., Armengaud P., Amtmann A. and Herzyk, P. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett, 573, 83-92.

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=15327980>

34 **BLAST.**

Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990) Basic local alignment search tool. J Mol Biol, 215, 403-10.

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=2231712>

35 **CRC.**

Zhaohui S. Qin (2006) Clustering microarray gene expression data using weighted Chinese restaurant process. Bioinformatics, 22(16), 1988-1997.

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=16766561>

36 **Filtering of gene expression data.**

Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E. S., Golub T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA, 96, 2907-2912.

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=10077610>

37 **CRC Tutorial.**

Steve Qin. CRC tutorial, 10/13/06

URL: <http://www.sph.umich.edu/csg/qin/CRC/tutor.pdf>

38 **Chinese restaurant clustering.**

Pitman J. (1996) Some Developments of the Blackwell-MacQueen Urn Scheme. IMS, Hayward, California.

39 **CRC Similarity measure.**

Medvedovic M. and Sivaganesan S. (2002) Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics, 18, 1194-1206.

PubMed <http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=pubmed&dopt=Abstract&list_uids=12217911>

40 **FreeType.**

Portions of this software are copyright © 2008 The FreeType Project (<http://www.freetype.org/>).

All rights reserved.

41 **TargetScan.**

<http://www.targetscan.org/>